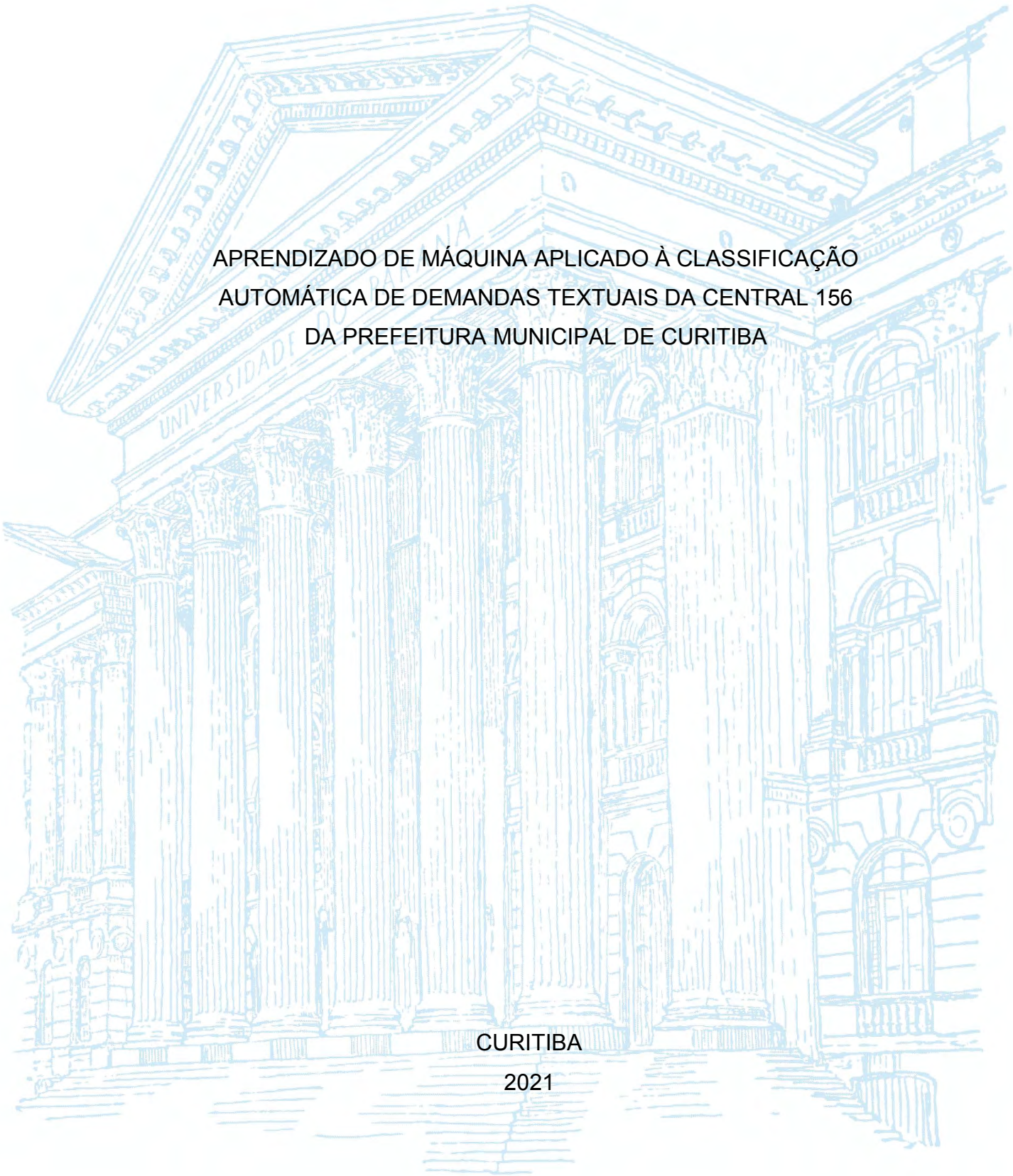
UNIVERSIDADE FEDERAL DO PARANÁ

LUCIMARA WONS



APRENDIZADO DE MÁQUINA APLICADO À CLASSIFICAÇÃO AUTOMÁTICA DE DEMANDAS TEXTUAIS DA CENTRAL 156 DA PREFEITURA MUNICIPAL DE CURITIBA

CURITIBA

2021

LUCIMARA WONS

APRENDIZADO DE MÁQUINA APLICADO À CLASSIFICAÇÃO AUTOMÁTICA DE DEMANDAS TEXTUAIS DA CENTRAL 156 DA PREFEITURA MUNICIPAL DE CURITIBA

Dissertação apresentada ao Curso de Pós-Graduação em Gestão da Informação do Setor de Ciências Sociais Aplicadas, Universidade Federal do Paraná, como requisito parcial à obtenção do título de Mestre em Gestão da Informação.

Orientador: Prof. Dr. Ricardo Mendes Junior

CURITIBA

2021

FICHA CATALOGRÁFICA ELABORADA PELA BIBLIOTECA DE CIÊNCIAS SOCIAIS APLICADAS – SIBI/UFPR COM DADOS FORNECIDOS PELO(A) AUTOR(A)
Bibliotecário: Maria Lidiane Herculano Graciosa – CRB 9/2018

Wons, Lucimara
Aprendizado de máquina aplicado à classificação automática de demandas textuais da central 156 da Prefeitura Municipal de Curitiba / Lucimara Wons. - 2021.
164 p.

Dissertação (Mestrado) - Universidade Federal do Paraná. Programa de Pós-Graduação em Gestão da Informação, do Setor de Ciências Sociais Aplicadas.

Orientador: Ricardo Mendes Junior.

Defesa: Curitiba, 2021.

1. Gestão da Informação. 2. Processamento de linguagem natural. 3. Processamento de texto. 4. Serviços Públicos. I. Universidade Federal do Paraná. Setor de Ciências Sociais Aplicadas. Programa de Pós-Graduação em Gestão da Informação. II. Mendes Junior, Ricardo. III. Título.

CDD 658.4038

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS SOCIAIS E APLICADAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO GESTÃO DA INFORMAÇÃO - 40001016058P1

ATA Nº112021

ATA DE SESSÃO PÚBLICA DE DEFESA DE MESTRADO PARA A OBTENÇÃO DO GRAU DE MESTRE EM GESTÃO DA INFORMAÇÃO

No dia vinte e dois de julho de dois mil e vinte e um às 10:00 horas, na sala https://conferenciaweb.rnp.br/webconf/ppggi-ufpr, Webconferência da RNP, foram instaladas as atividades pertinentes ao rito de defesa de dissertação da mestranda **LUCIMARA WONS**, intitulada: **APRENDIZADO DE MÁQUINA APLICADO À CLASSIFICAÇÃO AUTOMÁTICA DE DEMANDAS TEXTUAIS DA CENTRAL 156 DA PREFEITURA MUNICIPAL DE CURITIBA**, sob orientação do Prof. Dr. RICARDO MENDES JUNIOR. A Banca Examinadora, designada pelo Colegiado do Programa de Pós-Graduação em GESTÃO DA INFORMAÇÃO da Universidade Federal do Paraná, foi constituída pelos seguintes Membros: RICARDO MENDES JUNIOR (UNIVERSIDADE FEDERAL DO PARANÁ), DENISE FUKUMI TSUNODA (UNIVERSIDADE FEDERAL DO PARANÁ), ANDRÉA VASCONCELOS CARVALHO (UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE). A presidência iniciou os ritos definidos pelo Colegiado do Programa e, após exarados os pareceres dos membros do comitê examinador e da respectiva contra argumentação, ocorreu a leitura do parecer final da banca examinadora, que decidiu pela APROVAÇÃO. Este resultado deverá ser homologado pelo Colegiado do programa, mediante o atendimento de todas as indicações e correções solicitadas pela banca dentro dos prazos regimentais definidos pelo programa. A outorga de título de mestre está condicionada ao atendimento de todos os requisitos e prazos determinados no regimento do Programa de Pós-Graduação. Nada mais havendo a tratar a presidência deu por encerrada a sessão, da qual eu, RICARDO MENDES JUNIOR, lavrei a presente ata, que vai assinada por mim e pelos demais membros da Comissão Examinadora.

CURITIBA, 22 de Julho de 2021.

Assinatura Eletrônica
24/07/2021 08:30:30.0
RICARDO MENDES JUNIOR
Presidente da Banca Examinadora

Assinatura Eletrônica
23/07/2021 20:01:46.0
DENISE FUKUMI TSUNODA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
23/07/2021 19:04:51.0
ANDRÉA VASCONCELOS CARVALHO
Avaliador Externo (UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE)

MINISTÉRIO DA EDUCAÇÃO
SETOR DE CIÊNCIAS SOCIAIS E APLICADAS
UNIVERSIDADE FEDERAL DO PARANÁ
PRÓ-REITORIA DE PESQUISA E PÓS-GRADUAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO GESTÃO DA INFORMAÇÃO - 40001016058P1

TERMO DE APROVAÇÃO

Os membros da Banca Examinadora designada pelo Colegiado do Programa de Pós-Graduação em GESTÃO DA INFORMAÇÃO da Universidade Federal do Paraná foram convocados para realizar a arguição da dissertação de Mestrado de **LUCIMARA WONS** intitulada: **APRENDIZADO DE MÁQUINA APLICADO À CLASSIFICAÇÃO AUTOMÁTICA DE DEMANDAS TEXTUAIS DA CENTRAL 156 DA PREFEITURA MUNICIPAL DE CURITIBA**, sob orientação do Prof. Dr. RICARDO MENDES JUNIOR, que após terem inquirido a aluna e realizada a avaliação do trabalho, são de parecer pela sua APROVAÇÃO no rito de defesa.
A outorga do título de mestre está sujeita à homologação pelo colegiado, ao atendimento de todas as indicações e correções solicitadas pela banca e ao pleno atendimento das demandas regimentais do Programa de Pós-Graduação.

CURITIBA, 22 de Julho de 2021.

Assinatura Eletrônica
24/07/2021 08:30:30.0
RICARDO MENDES JUNIOR
Presidente da Banca Examinadora

Assinatura Eletrônica
23/07/2021 20:01:46.0
DENISE FUKUMI TSUNODA
Avaliador Interno (UNIVERSIDADE FEDERAL DO PARANÁ)

Assinatura Eletrônica
23/07/2021 19:04:51.0
ANDRÉA VASCONCELOS CARVALHO
Avaliador Externo (UNIVERSIDADE FEDERAL DO RIO GRANDE DO NORTE)

Dedico este trabalho ao meu esposo Leonardo, pela compreensão e força, e aos meus filhos Luciano e Letícia, pelo incentivo e pela paciência.

AGRADECIMENTOS

Agradeço a Deus pela dádiva da vida, por ser meu guia em todos os momentos e por me conceder fé e determinação para realizar mais este sonho.

Agradeço aos meus pais, Irineu e Tereza, pelo amor incondicional, pela educação que me proporcionaram e pelas palavras de conforto nas situações mais difíceis. Gratidão e orgulho por serem meus pais.

Agradeço imensamente ao meu esposo Leonardo e aos meus filhos Luciano e Letícia pela paciência e pelo apoio, principalmente nos momentos em que eu julgava carregar um fardo muito pesado, e por me fazerem acreditar que conseguiria.

Agradeço ao meu irmão, Cristiano, por demonstrar que nunca é tarde para alcançar nossos objetivos.

Agradeço ao meu orientador, Prof. Dr. Ricardo Mendes Junior, pelos ensinamentos, por ouvir minhas aflições e me acompanhar durante o mestrado. Meus sinceros respeito e consideração.

Agradeço aos membros da banca examinadora, Profª Drª Denise Fukumi Tsunoda e Profª Drª Andréa Vasconcelos Carvalho, por aceitarem o convite, pelas contribuições e pelas publicações indicadas, que auxiliaram e enriqueceram o desenvolvimento da pesquisa.

Agradeço aos professores do PPGGI, especialmente, à Profª Drª Maria do Carmo Freitas, pela oportunidade, carinho e confiança durante toda esta jornada, à Profª Drª Helena Nunes, pelas maravilhosas aulas de Metodologia Científica e ao Prof. Dr. Rodrigo Botelho Francisco, pelas produtivas aulas de Redação Científica.

Agradeço à Profª Drª Deborah Ribeiro Carvalho, da PUC-PR, pelas esclarecedoras aulas de inteligência artificial. Agradeço ao Prof. Dr. Walmes Marques Zeviani, do Departamento de Estatística, por disponibilizar, gratuitamente, vasto material sobre os temas desta pesquisa, os quais auxiliaram e facilitaram a elaboração do trabalho. Agradeço à bibliotecária Drª Janete Saldanha Estevão, por me ajudar a desvendar as bases de publicações científicas.

Agradeço aos colegas de mestrado, em especial, Leandra Ortigara, Ricardo Fujiwara, Karolayne Costa Rodrigues de Lima e Paulo Sergio da Conceição Moreira, pela parceria, pelo auxílio e aprendizado ao longo deste percurso.

Agradeço à prima de coração Eneida Regina Fabian Holzmann pela recomendação do PPGGI e pelas valiosas dicas.

Agradeço às minhas queridas amigas, Andréa Camboim, Ana Luiza Suplicy, Marcela Zanella e Luiza Ladika, pelo carinho e pelas boas vibrações.

Agradeço à secretária do PPGGI, Simone Batista, pela assistência, pelo profissionalismo e carinho dedicados.

Agradeço à Prefeitura Municipal de Curitiba, em especial à Srª Elvira Wos, ao Sr. Sozzi e ao colega Willian Belém, pela disponibilização de documentos em apoio à elaboração da pesquisa.

Agradeço também aos colegas da Central 156, Flávio dos Santos, Ozires Pereira de Oliveira e Carlos Fellipe Vidal da Cruz, por elucidarem dúvidas sobre o atendimento ao cidadão, realizado pela prefeitura por meio da Central 156.

RESUMO

Centrais de atendimento são utilizadas em muitos municípios brasileiros como meio de facilitar o relacionamento entre a prefeitura e os cidadãos, sendo responsáveis pelo contato com a população e registro de demandas referentes aos serviços públicos. O processo de atendimento nessas centrais produz grande quantidade de informações registradas em texto livre, que precisam de classificação manual para encaminhamento aos órgãos competentes. O objetivo da pesquisa é propor um modelo de classificação automática para as demandas textuais da Central 156 de Atendimento ao Cidadão da Prefeitura de Curitiba, utilizando-se algoritmos de aprendizado de máquina. Essa subárea da inteligência artificial, combinada ao processamento de linguagem natural, possibilita a classificação automática das demandas a partir da sua descrição, em aprimoramento ao processo de gestão da informação do atendimento ao cidadão da Central 156. A pesquisa, quanto ao propósito, caracteriza-se como descritiva, quanto à natureza como quali-quantitativa e quanto ao delineamento como documental e experimental. Analisa um *corpus* composto por 37.588 demandas em texto livre, coletadas do Portal de Dados Abertos da Prefeitura de Curitiba e obtidas a partir de amostragem aleatória e *undersampling*. A metodologia experimental da pesquisa é orientada pelo método CRISP-DM e as demandas estão distribuídas segundo oito órgãos da prefeitura, que totalizam 98% do total das demandas de 2019. O *corpus* é submetido ao processamento de linguagem natural, com tratamento para o idioma português, e as características resultantes são representadas no modelo espaço vetorial como unigramas e bigramas, utilizando a ponderação de termos TF-IDF. Além de remoção de *stopwords* e conflação por *stemming*, são aplicados valores limites para redução de esparsidade e dimensionalidade do modelo. Os resultados indicam bons níveis de concordância entre a classificação realizada pelos atendentes e a obtida nos experimentos. Com o algoritmo Naïve Bayes Multinomial, para unigramas, o coeficiente de Kappa atinge 0,90 e a taxa de acerto 91,3% com o tempo de processamento de 6 segundos. Como principal resultado da pesquisa tem-se um modelo para classificação automática das demandas em três estágios, iniciando por órgão, depois por assunto e, então, por subdivisão. Nesta aplic33ação, a classe FAS foi a que apresentou desempenho superior e a SMDT o mais baixo, evidenciando que os termos que aparecem somente em determinada classe influenciam positivamente os coeficientes de Kappa e as taxas de acerto obtidas. Uma contribuição relevante da pesquisa é o seu potencial uso na classificação das demandas da Central 156, em auxílio aos atendentes, bem como na classificação de demandas com entrada em outros canais da Prefeitura de Curitiba, como a Lei de Acesso à Informação e o Fala Curitiba.

Palavras-chave: Processamento de linguagem natural. Classificação de textos. Relacionamento governo-cidadão. Serviços públicos. Gestão da informação.

ABSTRACT

Call centers are used in many Brazilian municipalities as a way of facilitating the relationship between city hall and citizens, being responsible for the contact of the population and public service demands registry. The service process of these centers creates great volume of recorded free text which need to be manually assorted in order to be forwarded to the adequate government department. The objective of this research is to propose an automatic classification model for the text demands of the Central 156 of Citizen Service of Curitiba's City Hall through machine learning algorithms. This subarea of artificial intelligence, combined with natural language processing, allows an automatic classification of the demands from their descriptions in improvement on the information management process of the citizen service of the Central 156. The research in regards of purpose is classified as descriptive, in regards of nature as qualitative-quantitative and in regards of procedure as documental and experimental. It analyses a corpus of 37,588 demands in free text gathered from random sampling and undersampling of the Curitiba's City Hall Open Data Portal. This research's experimental methodology is guided by the CRISP-DM method and the demands are distributed according to eight departments of the city hall which totalize 98% of all demands of 2019. The corpus was submitted to the natural language processing with treatment for the Portuguese language and the resulting characteristics are presented in a space vector template as unigrams and bigrams using the TF-IDF term weighting. Other than stopword removal and stemming conflation, limit values are applied in order to reduce the sparsity and dimensionality of the model. The results indicate good levels of concordance between the manual classification and the classification obtained in the experiments. With the Naïve Bayes Multinomial algorithm for unigrams the Kappa coefficient achieved is 0.9 with a success rate of 91.3% and a processing time of 6 seconds. The main goal of this research is a model for automatic classification of demands in three stages, starting with department, then subject and finally subdivision. In this instance, the FAS class produced superior performance and the SMDT class produced the lowest performance, showing that terms that only appear in determined classes positively influenced the Kappa coefficients and obtained success rates. A relevant contribution of this research is its potential use in the classification of Central 156 demands in aid of clerks, as well as the classification of demands that come from other sources of the Curitiba's City Hall, such as the Information Access Law and the Fala Curitiba.

Keywords: Natural language processing. Text classification. Citizen government relationship. Public services. Information Management.

LISTA DE FIGURAS

FIGURA 1 – SUBÁREAS DA INTELIGÊNCIA ARTIFICIAL 22
FIGURA 2 – FASES DO USO ESTRATÉGICO DA INFORMAÇÃO: CRIAÇÃO DE SIGNIFICADO, CONSTRUÇÃO DE CONHECIMENTO E TOMADA DE DECISÃO 32
FIGURA 3 – DIAGRAMA MACRO DAS DEFINIÇÕES INVESTIGADAS E PRODUTOS INFORMACIONAIS ELABORADOS 37
FIGURA 4 – SÍNTESE DOS NÍVEIS DE AMADURECIMENTO DO RELACIONAMENTO ENTRE GOVERNO E CIDADÃO 44
FIGURA 5 – DIAGRAMA REFERENTE AO TÓPICO APRENDIZADO DE MÁQUINA 62
FIGURA 6 – DIAGRAMA REFERENTE AO TÓPICO PROCESSAMENTO DE LINGUAGEM NATURAL 66
FIGURA 7 – GRÁFICO DE LUHN RELACIONANDO A EXPRESSIVIDADE E FREQUÊNCIA DE OCORRÊNCIA DAS PALAVRAS 70
FIGURA 8 – DIAGRAMA REFERENTE AO TÓPICO REPRESENTAÇÃO DE DOCUMENTOS 71
FIGURA 9 – DIAGRAMA REFERENTE AO TÓPICO MINERAÇÃO DE TEXTOS ...74
FIGURA 10 – SÍNTESE DA CARACTERIZAÇÃO DA PESQUISA 81
FIGURA 11 – FASES DA METODOLOGIA CRISP-DM 87
FIGURA 12 – FUNCIONALIDADES DO SIAC-156 93
FIGURA 13 – TELA DE REGISTRO DA DEMANDA NO SIAC-156, COM DESTAQUE PARA O CAMPO DESCRIÇÃO.....94
FIGURA 14 – EXEMPLIFICAÇÃO DO PRÉ-PROCESSAMENTO DO TEXTO DA DEMANDA DO 156 102
FIGURA 15 – FLUXO DO PROCESSO SIMPLIFICADO DE GESTÃO DA INFORMAÇÃO DO ATENDIMENTO AO CIDADÃO DA PMC, VIA CENTRAL 156 109
FIGURA 16 – MATRIZES DE CONFUSÃO OBTIDAS NA CLASSIFICAÇÃO COM UNIGRAMAS – *DATASET* MAIOR 112
FIGURA 17 – MATRIZES DE CONFUSÃO OBTIDAS NA CLASSIFICAÇÃO COM UNIGRAMAS – *DATASET* MENOR.....113

FIGURA 18 – MATRIZES DE CONFUSÃO OBTIDAS NA CLASSIFICAÇÃO COM BIGRAMAS – *DATASET* MAIOR 115
FIGURA 19 – MATRIZES DE CONFUSÃO OBTIDAS NA CLASSIFICAÇÃO COM BIGRAMAS – *DATASET* MENOR 116
FIGURA 20 – FLUXO SIMPLIFICADO DE ATENDIMENTO AO CIDADÃO - CANAIS *CHAT* HUMANO E TELEFONE INCLUINDO A AUTOMATIZAÇÃO DA CLASSIFICAÇÃO DAS DEMANDAS 119
FIGURA 21 – ETAPAS DO MODELO PROPOSTO DE CLASSIFICAÇÃO AUTOMÁTICA DAS DEMANDAS EM TRÊS ESTÁGIOS 120

LISTA DE GRÁFICOS

GRÁFICO 1 – EVOLUÇÃO DAS PUBLICAÇÕES PARA OS TERMOS ENVOLVENDO APRENDIZADO DE MÁQUINA, PROCESSAMENTO DE LINGUAGEM NATURAL E CLASSIFICAÇÃO DE TEXTOS, NO PERÍODO ENTRE 1965 E 2019 27
GRÁFICO 2 – NÚMERO DE PAÍSES AGRUPADOS SEGUNDO ÍNDICE DE DESENVOLVIMENTO DE E-GOV (EGDI) EM 2018 41
GRÁFICO 3 – ATENDIMENTOS REALIZADOS PELA CENTRAL 156 DE CURITIBA, SEGUNDO CANAIS DE ATENDIMENTO, NO MÊS DE DEZEMBRO DE 2019 92

LISTA DE QUADROS

QUADRO 1 – PESQUISAS DO PPGGI COM TEMA VINCULADO À CLASSIFICAÇÃO DE TEXTOS 28
QUADRO 2 – ALGUNS DOS TRABALHOS RELACIONADOS À PESQUISA E AOS CONCEITOS ABORDADOS 75
QUADRO 3 – SÍNTESE DO ALINHAMENTO ENTRE OBJETIVOS ESPECÍFICOS, CONCEITOS E AUTORES 85
QUADRO 4 – DICIONÁRIO DOS DADOS ABERTOS DA CENTRAL 156 DE CURITIBA 96
QUADRO 5 – EXEMPLO DE DEMANDA REGISTRADA PELA CENTRAL 156 EM 2019 97
QUADRO 6 – RECURSOS UTILIZADOS NO DESENVOLVIMENTO DO MÉTODO, OBJETOS E FUNCIONALIDADES ENVOLVIDOS 107

LISTA DE TABELAS

TABELA 1 – REPRESENTAÇÃO DE DOCUMENTOS NO MODELO *BAG-OF-WORDS* 67
TABELA 2 – EXPRESSÕES DE BUSCA, PORTAL E BASES DE DADOS UTILIZADOS NA PESQUISA DE PUBLICAÇÕES CIENTÍFICAS 82
TABELA 3 – EXPRESSÕES DE BUSCA, BASE DE DADOS E PORTAL BRASILEIROS UTILIZADOS NA PESQUISA DE TESES E DISSERTAÇÕES 83
TABELA 4 – ELIMINAÇÃO DE REGISTROS INCONSISTENTES OU NÃO RELEVANTES 98
TABELA 5 – DISTRIBUIÇÃO DAS DEMANDAS DO 156 SEGUNDO ÓRGÃOS DA PMC EM 2019 99
TABELA 6 – DISTRIBUIÇÃO DOS EXEMPLOS ENTRE AS OITO CLASSES 100
TABELA 7 – DIMENSÃO DAS DTMs COM UNIGRAMAS E BIGRAMAS, ANTES E APÓS A REDUÇÃO DE ESPARSIDADE – *CORPUS* DA CLASSIFICAÇÃO POR ÓRGÃO 110
TABELA 8 – RESULTADOS OBTIDOS NA CLASSIFICAÇÃO COM UNIGRAMAS 111
TABELA 9 – RESULTADOS OBTIDOS NA CLASSIFICAÇÃO COM BIGRAMAS .. 114

LISTA DE ABREVIATURAS OU SIGLAS

ADM - Órgãos da administração
AM - Aprendizado de Máquina
ANS - Acordo de Nível de Serviço
AP - Aprendizado Profundo
BDTD - Biblioteca Digital Brasileira de Teses e Dissertações
BOW - *Bag-of-words* ou saco de palavras
CAFe - Comunidade Acadêmica Federada
CiRM - C*itizen Relationship Management* ou Gestão de Relacionamento com o Cidadão
CPD - Centro de Processamento de Dados
CRM - *Customer Relationship Management* ou Gestão de Relacionamento com o Cliente
DTM - *Document term matrix* ou matriz de termos e documentos
EGDI - E-*Government Development Index* ou Índice de Desenvolvimento em Governo Eletrônico
e-gov - E*letronic government* ou governo eletrônico
FAS - Fundação de Ação Social
GI - Gestão da Informação
IA - Inteligência Artificial
IBCTI - Instituto Brasileiro de Informação em Ciência e Tecnologia
IBM - *International Business Machines* (empresa)
IBk - Instance-Based k-Nearest Neighbor ou Baseado em instância pelo vizinho mais próximo
IDF - *Inverse document frequency* ou frequência inversa do documento
ICI - Instituto das Cidades Inteligentes
IPPUC - Instituto de Pesquisa e Planejamento Urbano de Curitiba
IPTU - Imposto Predial e Territorial Urbano
ISS - Imposto sobre serviço
k-NN - *k-nearest neighbors* ou k-vizinhos mais próximos
NLP - *Natural Language Processing*
NLU - *Natural Language Understanding*
PLN - Processamento de Linguagem Natural

PMC - Prefeitura Municipal de Curitiba
RSO - Responsável pelo serviço no órgão
SIAC-156 - Sistema Integrado de Atendimento ao Cidadão-156
SMDT - Secretaria Municipal da Defesa Social e Trânsito
SMOP - Secretaria Municipal de Obras Públicas
SGM - Secretaria do Governo Municipal
SMMA - Secretaria Municipal do Meio Ambiente
SMS - Secretaria Municipal da Saúde
SMU - Secretaria Municipal do Urbanismo
TDM - *Term document matrix* ou matriz de documentos e termos
TICs - Tecnologias da Informação e Comunicação
TF - *Term frequency* ou frequência do termo
TF-IDF - *Term frequency - inverse document frequency* ou frequência do termo - frequência inversa no documento
UFPR - Universidade Federal do Paraná
URBS - Urbanização de Curitiba
WEKA - *Waikato Environment for Knowledge Analysis*

SUMÁRIO

1 INTRODUÇÃO 20
1.1 CONTEXTUALIZAÇÃO DO PROBLEMA 24
1.2 OBJETIVOS 25
1.2.1 Objetivo geral 25
1.2.2 Objetivos específicos 25
1.3 JUSTIFICATIVA 26
1.4 DELIMITAÇÃO DA PESQUISA 30
1.5 ESTRUTURA DA DISSERTAÇÃO 30
2 REVISÃO DE LITERATURA 31
2.1 GESTÃO DA INFORMAÇÃO NA ADMINISTRAÇÃO PÚBLICA 31
2.2 RELACIONAMENTO GOVERNO-CIDADÃO POR MEIO DA TECNOLOGIA DA INFORMAÇÃO 37
2.2.1 Níveis de amadurecimento do relacionamento governo-cidadão 38
2.3 APRENDIZADO DE MÁQUINA 45
2.3.1 Abordagens do aprendizado de máquina 46
2.3.2 Aprendizado supervisionado 50
2.3.3 Classificação de textos 54
2.3.4 Algoritmos de classificação de textos 57
2.4 PROCESSAMENTO DE LINGUAGEM NATURAL 63
2.5 REPRESENTAÇÃO DE DOCUMENTOS 66
2.5.1 Ponderação de termos 68
2.5.2 Redução de dimensionalidade 69
2.6 MINERAÇÃO DE TEXTOS 72
2.7 TRABALHOS RELACIONADOS 74
3 ENCAMINHAMENTOS METODOLÓGICOS 80
3.1 CARACTERIZAÇÃO DA PESQUISA 80
3.2 MATERIAIS E MÉTODOS 81
3.2.1 Levantamento bibliográfico 81
3.2.2 Pesquisa documental 86
3.2.3 Pesquisa experimental 86
4 DESENVOLVIMENTO DO MÉTODO CRISP-DM 89
4.1 COMPREENSÃO DO NEGÓCIO 89

4.1.1 Ambiente da pesquisa 89
4.1.2 Recursos utilizados para desenvolvimento do método CRISP-DM 95
4.2 COMPREENSÃO DOS DADOS 95
4.3 PREPARAÇÃO DOS DADOS 101
4.4 MODELAGEM 104
4.5 AVALIAÇÃO 105
4.6 SÍNTESE DOS RECURSOS UTILIZADOS 107
5 RESULTADOS 108
5.1 PROCESSO DE GESTÃO DA INFORMAÇÃO DO ATENDIMENTO AO CIDADÃO DA CENTRAL 156 108
5.2 RESULTADOS DO PROCESSAMENTO DE LINGUAGEM NATURAL E DA REPRESENTAÇÃO DOS DOCUMENTOS 109
5.3 RESULTADOS DA APLICAÇÃO DOS ALGORITMOS DE APRENDIZADO DE MÁQUINA 110
5.3.1 Resultados da classificação com unigramas 110
5.3.2 Resultados da classificação com bigramas 114
5.3.3 Comparação dos resultados 116
5.4 MODELO DE CLASSIFICAÇÃO PROPOSTO 118
6 CONSIDERAÇÕES FINAIS 121
6.1 LIMITAÇÕES DA PESQUISA 124
6.2 RECOMENDAÇÕES PARA TRABALHOS FUTUROS 124
REFERÊNCIAS 125
APÊNDICE 1 – FLUXO SIMPLIFICADO DE ATENDIMENTO AO CIDADÃO POR MEIO DOS CANAIS TELEFONE E *CHAT* HUMANO 136
APÊNDICE 2 – FLUXO SIMPLIFICADO DE ATENDIMENTO AO CIDADÃO POR MEIO DO CANAL PORTAL *WEB* 137
APÊNDICE 3 – FLUXO SIMPLIFICADO DE ATENDIMENTO AO CIDADÃO POR MEIO DO CANAL *CHAT* AUTOMATIZADO 138
APÊNDICE 4 – FLUXO SIMPLIFICADO DE ATENDIMENTO AO CIDADÃO POR MEIO DO CANAL APP CURITIBA 156 - *MOBILE* 139
APÊNDICE 5 – CÁLCULO PARA OBTENÇÃO DOS EXEMPLOS, UTILIZANDO *UNDERSAMPLING* COM REGISTROS DA SMMA 141
APÊNDICE 6 – CÁLCULO PARA OBTENÇÃO DOS EXEMPLOS, UTILIZANDO *UNDERSAMPLING* PARA AS DEMAIS CLASSES DO ESTUDO 144

APÊNDICE 7 – LISTA DE STOPWORDS UTILIZADAS 151
APÊNDICE 8 – DETALHAMENTO DO PROCESSO DE GESTÃO DA INFORMAÇÃO DA PMC, REALIZADO VIA CENTRAL 156 152
APÊNDICE 9 – RESULTADOS OBTIDOS NA CLASSIFICAÇÃO COM UNIGRAMAS – *DATASET* MAIOR 154
APÊNDICE 10 – RESULTADOS OBTIDOS NA CLASSIFICAÇÃO COM UNIGRAMAS – *DATASET* MENOR 155
APÊNDICE 11 – RESULTADOS OBTIDOS NA CLASSIFICAÇÃO COM BIGRAMAS – *DATASET* MAIOR 157
APÊNDICE 12 – RESULTADOS OBTIDOS NA CLASSIFICAÇÃO COM BIGRAMAS – *DATASET* MENOR 158
APÊNDICE 13 – AUTORIZAÇÃO PARA USO DOS DADOS, DOCUMENTOS E DO NOME DA PREFEITURA NA PESQUISA 160

1 INTRODUÇÃO

A informação é caracterizada como algo significativo para os indivíduos e, geralmente, interpretada a partir de uma representação simbólica contida em dados (SETZER, 1999, p. 2). A percepção da informação recebida, associada às aprendizagens anteriores e ao raciocínio, resulta na aquisição do conhecimento (SORDI, 2008, p. 9-12) que, no contexto organizacional, pode ser traduzido em maior eficiência e competitividade.

Dados podem ser estruturados, como registros padronizados em bancos de dados, e não estruturados (FELDMAN; SANGER, 2007, p. 1), compreendendo documentos, relatórios, mensagens e vídeos, os quais correspondem a 80% dos dados produzidos no mundo (SCHNEIDER, 2016, p. 2). Mesmo que constituam uma fonte significativa de oportunidades de negócios para as organizações, tem sido um desafio transformar dados não estruturados em informação útil (CASTRO; FERRARI, 2016, p. 34; LAROSE; LAROSE, 2014, p. xi).

Considerando que o volume de dados produzidos no mundo era da ordem de 2,5 bilhões de gigabytes diários em 2016 (SCHNEIDER, 2016, p. 2), a gestão da informação (GI) se faz um recurso indispensável às organizações, visto que possibilita gestão dos processos e sistemas organizacionais voltados à aquisição, criação, organização, distribuição e uso de informações (CHOO, 2020).

Como em outros domínios organizacionais, a gestão da informação é realizada a partir de modelos que segmentam o processo em tarefas interdependentes, ou etapas, e que seguem uma sequência lógica de execução. Para McGee e Prusak (1994, p. 108) essas etapas contemplam identificação das necessidades de informação, coleta, armazenamento, tratamento e desenvolvimento de produtos e serviços, além da distribuição e do uso da informação. Com a adoção da gestão da informação, os colaboradores dinamizam a realização de suas atividades, fazendo com que a organização opere de forma mais produtiva, competitiva e rentável (DETLOR, 2010, p. 103).

Considerando o grande volume de dados existente nas organizações, um dos meios de aprimorar a gestão da informação é a utilização das técnicas de inteligência artificial (IA), a qual diz respeito à programação ou treinamento do computador para realizar tarefas até então reservadas à inteligência humana, tais como recomendar filmes e responder questões técnicas (MEHR; ASH; FELLOW, 2017, p. 1).

Na administração pública, uma parte significativa do volume de dados e informações gerados decorre da interação com os cidadãos, no que se refere às demandas por serviços públicos. Nesse âmbito, a IA tem o potencial de causar um grande impacto na maneira como os cidadãos vivenciam e interagem com o governo. Embora não seja uma solução para os problemas, a IA é uma ferramenta poderosa para aumentar a eficiência do governo. A implementação e o uso de IA nos serviços aos cidadãos também podem se tornar um indicador de como o setor público pode alavancar outras ferramentas digitais emergentes, abrindo caminho para envolvimento e *feedback* dos cidadãos quanto a essas tecnologias (MEHR; ASH; FELLOW, 2017, p. 15).

O Governo do Canadá, por exemplo, numa de suas recentes diretrizes, propõe o uso da IA para auxiliar no atendimento ao cidadão (KUZIEMSKI; MISURACA, 2020, p. 5). A diretriz canadense estabelece que pesquisas inteligentes possam identificar padrões nos dados fornecidos, com o objetivo de entender melhor o que os cidadãos desejam ao acessar os serviços governamentais; que programas de computador, ou *chatbots*, filtrem perguntas de rotina dos cidadãos, permitindo que os funcionários possam se concentrar em casos complexos e; que exista suporte automatizado de decisão, aumentando a qualidade dos serviços e reduzindo os tempos de espera.

A abordagem proposta por Androutsopoulou *et al.* (2018), na Grécia, utiliza *chatbots* com o propósito de desenvolver um novo canal de comunicação entre governos e cidadãos. A partir de dados existentes como leis, diretrizes, dados provenientes dos sistemas de órgãos governamentais e das mídias sociais, a ferramenta permite interação, em linguagem coloquial, tanto para busca de informações quanto para realização de transações.

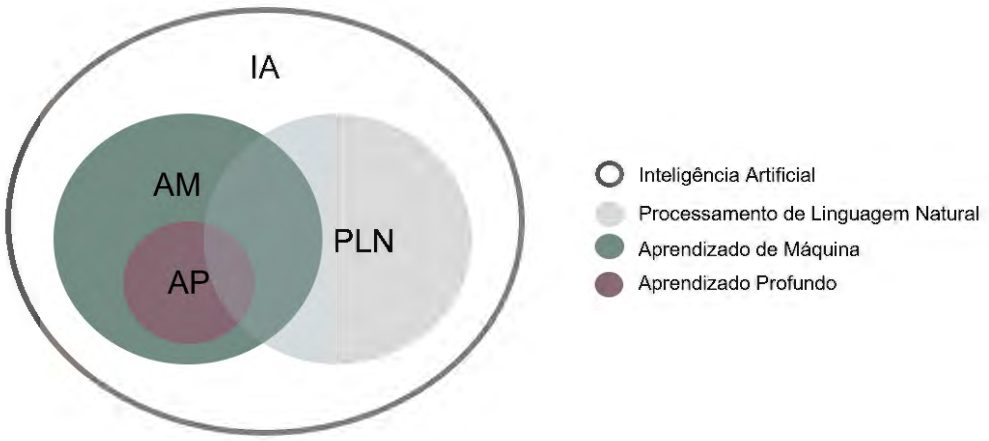
No Brasil, encontra-se em tramitação, pelo Senado Federal, o Projeto de Lei n° 5.691/2019 que institui a Política Nacional de Inteligência Artificial (BRASIL, 2019). Nesse projeto, a IA é considerada um recurso potencial para impulsionar o crescimento e o aumento da produtividade das organizações e para otimizar o tempo das pessoas. A implementação da IA, segundo o projeto, deve envolver esforços do governo, das indústrias e universidades. Dentre suas diretrizes estão a melhoria da qualidade e da eficiência dos serviços prestados pelo governo, o estímulo a investimentos para desenvolvimento da IA e para realização de atividades voltadas à pesquisa e inovação. Vislumbra-se como melhorias, por exemplo, agilização da

prestação de serviços públicos, em se tratando do fornecimento de informações, bem como esclarecimento ao cidadão acerca dos processos e dados necessários para realização de determinados serviços.

Processos de automação são muito bem-vindos em trabalhos que envolvem monitoramento contínuo, que consomem tempo e que são tediosos para os humanos (WITTEN *et al*., 2017, p. 28). Na esfera governamental, a combinação dos algoritmos de IA, essencialmente de aprendizado de máquina, às grandes quantidades de dados existentes, pode melhorar a operacionalização dos processos, viabilizando modelos de prestação de serviços proativos e auxiliando na execução de tarefas rotineiras e repetitivas (KUZIEMSKI; MISURACA, 2020, p. 3). Mehr, Ash e Fellow (2017, p. 1-2) afirmam que é necessário modernizar o atendimento ao cidadão, uma vez que este se beneficia de sistemas do Século XXI em seu cotidiano, porém, ao interagir com o governo, se depara com sistemas antigos, do Século XX.

Além do aprendizado de máquina, outras duas subáreas da IA se destacam no relacionamento entre governo e cidadão: o processamento de linguagem natural (PLN) e o aprendizado profundo, conforme apresentado na Figura 1.

FIGURA 1 – SUBÁREAS DA INTELIGÊNCIA ARTIFICIAL



FONTE: AthenaTech LLC (2019)

O aprendizado de máquina busca desenvolver sistemas capazes de aprender de forma automática (LOPES; SANTOS; PINHEIRO, 2014, p. 3; MITCHELL, 1997, p. 1; MONARD; BARANAUSKAS, 2003, p. 39) e o processamento de linguagem natural possibilita que computadores sejam empregados para manipular e entender textos em linguagem natural (CHOWDHURY, 2003, p. 51). O aprendizado profundo, ou *deep*

learning, no inglês, compreende redes neurais artificiais complexas que extraem representações de padrões implícitos em conjuntos de dados (WITTEN *et al*., 2017, p. 206).

No relacionamento entre governo e cidadãos, o aprendizado de máquina pode auxiliar na redução de encargos administrativos e na resolução de problemas de alocação de recursos, uma vez que os serviços aos cidadãos compreendem questões como pesquisar e verificar o trâmite de documentos e elaborar documentos solicitados (MEHR; ASH; FELLOW*,* 2017, p. 1-2). Os autores afirmam que, com a utilização dessa tecnologia, o governo pode tornar-se mais eficiente, pois, ao atuar em atividades antes realizadas pelos funcionários, estes são liberados para oferecer um atendimento melhor, aumentando a satisfação dos cidadãos em relação aos serviços prestados.

Uma possibilidade de aplicação do aprendizado de máquina no relacionamento entre governo e cidadãos é nas centrais de atendimento, existentes na maioria das cidades. No Brasil, essas centrais atendem pelo telefone 156, sendo chamadas de Centrais 156, número regulamentado pela Agência Nacional de Telecomunicações como de utilidade pública e específico para atendimento referente aos serviços municipais (MINISTÉRIO DAS COMUNICAÇÕES, 2015).

Centrais de atendimento ao cidadão são utilizadas por muitos municípios brasileiros como meio de estreitar e facilitar o relacionamento entre a prefeitura e os cidadãos. Essas centrais são responsáveis pelo contato com a população e registro de demandas, incluindo solicitações, reclamações, sugestões e denúncias alusivas aos serviços públicos. O processo de atendimento nessas centrais produz grande quantidade de informações registradas, na maioria dos casos, em formato texto. O aprendizado de máquina pode ser utilizado para o tratamento destes textos possibilitando, por exemplo, sua classificação, a busca por fontes de informação e a busca por respostas automáticas.

Assim, esta pesquisa visa contribuir com a melhoria do relacionamento entre governo a cidadãos via centrais de atendimento, estudando a aplicação de aprendizado de máquina para classificação dos textos referentes às demandas dos cidadãos, com entrada na Central 156 de Curitiba.

1.1 CONTEXTUALIZAÇÃO DO PROBLEMA

Para investigar o problema indicado acima, esta pesquisa utiliza a Central 156 da Prefeitura de Curitiba (PMC) como ambiente de estudo. A Central 156 disponibiliza cinco canais para registro das demandas: o número telefônico 156, o portal *web*, o *chat* humano, o *chat* automatizado, ou robô, e o aplicativo *mobile* Curitiba 156. Em 2019, conforme os dados abertos disponibilizados pela prefeitura, a Central 156 registrou mais de 316 mil demandas dos cidadãos (PMC, 2019a).

Por meio do processo de gestão da informação da Central 156 da PMC, definiu-se que, nos canais telefone e *chat* humano, as demandas são coletadas e armazenadas em texto livre, a partir do relato dos cidadãos, sendo classificadas manualmente pelos atendentes da central. Nessa etapa, cada demanda recebe dois atributos, o assunto e a subdivisão, necessários para associá-la a um determinado serviço da prefeitura. Essa classificação permite que a demanda seja encaminhada ao órgão responsável pela análise, realização do serviço, se for o caso, e resposta para o cidadão.

O tempo médio de resposta das demandas foi de cinco dias[1]. Esse tempo compreende o número de dias entre a abertura do protocolo pelo cidadão e o lançamento da resposta no sistema. A resposta geralmente apresenta um parecer sobre a solicitação e informa a realização do serviço, a previsão ou a justificativa para a não realização. Há ainda a resposta administrativa, utilizada quando a natureza do serviço não envolve uma execução ou quando concerne a elogios ou reclamações.

Além do tempo despendido pelo órgão para dar retorno ao cidadão, outro fator que influi no tempo de resposta é a tramitação correta para o órgão responsável. Quando isso não ocorre, é necessário que o responsável pelo serviço no órgão solicite à Central 156 novo encaminhamento da demanda. Em 2019, 3,5% dos protocolos[1] apresentaram *status* “alterar assunto/responsável”, indicando a ocorrência dessa situação.

Dentre as técnicas de aprendizado de máquina a classificação permite, a partir de um grupo de elementos com classes ou rótulos conhecidos a *priori*, treinar e ajustar um modelo computacional para efetuar a classificação automaticamente

---

[1] Disponível em: http://dadosabertos.c3sl.ufpr.br/curitiba/156/. Análise amostral realizada pela autora com protocolos referentes aos meses de março, julho e novembro de 2019. Acesso em: 15 fev. 2020.

(CASTRO; FERRARI, 2016, p. 41). O aprendizado de máquina, associado ao processamento de linguagem natural, possibilita classificar textos livres automaticamente. A utilização da classificação com aprendizado de máquina para o tratamento das descrições das demandas na Central 156 abre campo para uma série de aplicações como, por exemplo, apresentar ao atendente uma lista dos serviços mais condizentes à demanda; efetuar automaticamente a tramitação para o órgão responsável e; até mesmo, automatizar a resposta para o cidadão. Essas funcionalidades são meios de aprimorar o processo de gestão da informação da Central 156, ao tratar o grande volume de dados existente e possibilitar a redução do tempo de resposta, corroborando o atendimento realizado.

Diante do exposto, considera-se possível que a aplicação das técnicas de aprendizado de máquina na classificação das demandas por serviços municipais possa trazer benefícios para os cidadãos, atendentes da Central 156 e responsáveis pelos serviços nos órgãos da prefeitura. Assim sendo, insere-se a questão da presente dissertação: **como classificar as demandas da Central 156 de Curitiba utilizando o aprendizado de máquina?**

1.2 OBJETIVOS

A partir da contextualização e do problema de pesquisa, foram traçados o objetivo geral e os objetivos específicos para este trabalho.

1.2.1 Objetivo geral

Propor um modelo de classificação automática para as demandas textuais da Central 156 de Atendimento ao Cidadão da Prefeitura de Curitiba, por meio de algoritmos de aprendizado de máquina.

O objetivo geral foi desdobrado em objetivos específicos, conforme segue.

1.2.2 Objetivos específicos

Os objetivos específicos que auxiliam a elaboração da pesquisa são:

a) descrever o processo de gestão da informação do atendimento ao cidadão, realizado via Central 156;

b) submeter os textos das demandas ao processamento de linguagem natural, representando-os no modelo espaço vetorial;
c) aplicar algoritmos de aprendizado de máquina para classificação das demandas por órgão.

1.3 JUSTIFICATIVA

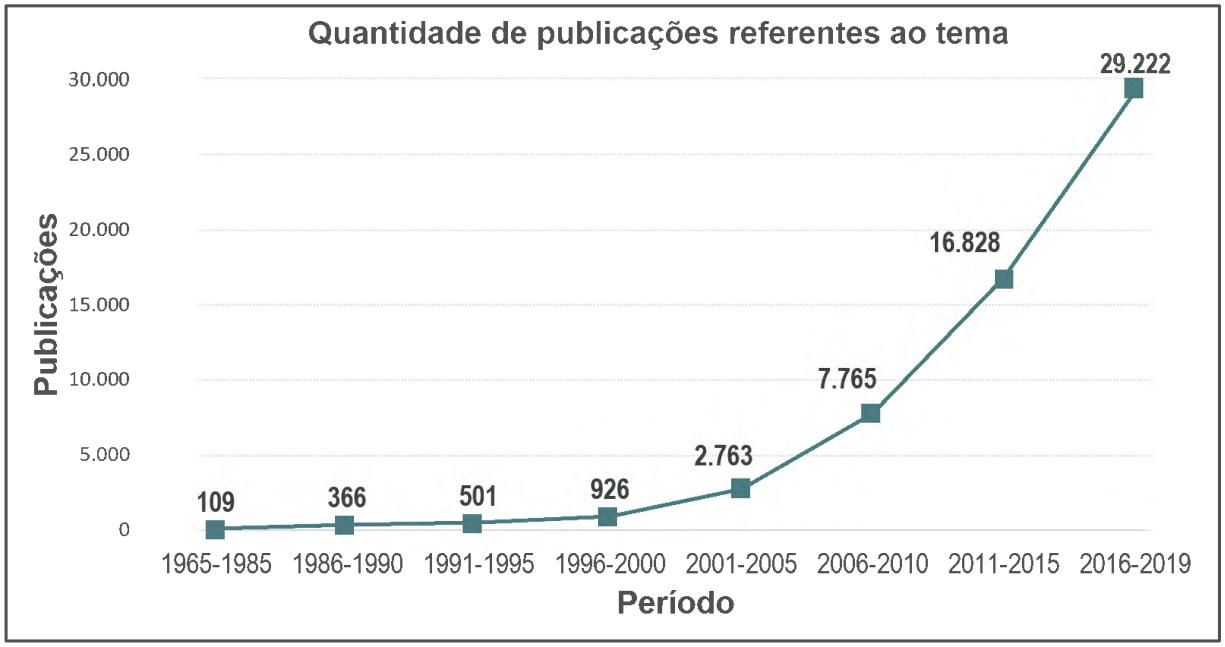
O tema central da pesquisa, envolvendo “aprendizado de máquina”, “processamento de linguagem natural” e “classificação de textos”, tem sido objeto de investigação de muitos pesquisadores, principalmente nos últimos 20 anos, conforme resultados do levantamento bibliográfico realizado em julho de 2019, início da elaboração deste estudo.

O levantamento foi realizado para os idiomas português e inglês, considerando-se o período entre 1965 e 2019, e ocorreu em duas etapas. Na primeira, foram utilizados os termos “aprendizado de máquina”, “sistemas inteligentes”, "inteligência computacional", “classificação de textos”, “categorização de textos”, “mineração de textos”, “processamento de textos” e “processamento de linguagem natural”, resultando em 58.480 publicações, conforme apresentado no Gráfico 1.

Observa-se que há um crescimento exponencial no número de publicações ao longo dos 54 anos examinados. O maior aumento ocorreu entre 1985 e 1990, quando o número mais que triplicou. Nos quinquênios 2000-2005 e 2005-2010 a alta persiste, atingindo quase o triplo de publicações. Entre 2010 e 2015, o número foi maior que o dobro, e quase o dobro para o último período examinado, entre 2015 e 2019. Esses números refletem a importância do tema deste trabalho.

Na segunda etapa, e com vistas a focar a busca em serviços públicos, foram adicionados os termos “serviços públicos”, “serviços municipais”, “serviços emergenciais", "serviços ao cidadão", "serviços governamentais” e “atendimento ao cidadão”, resultando em apenas 504 publicações. Esse total representa menos de 1% dos trabalhos obtidos para o tema, evidenciando que há espaço para estudos que contemplem a aplicação das técnicas citadas, no contexto dos serviços públicos.

GRÁFICO 1 – EVOLUÇÃO DAS PUBLICAÇÕES PARA OS TERMOS ENVOLVENDO APRENDIZADO DE MÁQUINA, PROCESSAMENTO DE LINGUAGEM NATURAL E CLASSIFICAÇÃO DE TEXTOS, NO PERÍODO ENTRE 1965 E 2019



FONTE: A autora (2019)

Posteriormente, foi realizado levantamento, também em duas etapas e com os mesmos termos de busca, com vistas à obtenção dos totais de teses e dissertações publicadas em português, no mesmo período. Foram recuperadas 524 publicações, sendo 21 com foco em serviços públicos. Novamente, é possível observar a escassez de trabalhos acerca do tema, em se tratando de serviços públicos, denotando a pertinência do estudo ora apresentado. Dessa maneira, os resultados obtidos nas buscas respaldam a justificativa acadêmica da dissertação.

As expressões de busca, assim como todos os termos utilizados e os achados do levantamento realizado, são expostos no Capítulo 3 – Encaminhamentos metodológicos, tópico 3.2.1 – Levantamento bibliográfico.

A pesquisa mostra-se alinhada ao Programa de Pós-Graduação em Gestão da Informação (PPGGI) da Universidade Federal do Paraná (UFPR) ao apresentar estudo quanto à aplicação do aprendizado de máquina para transformar dados textuais em informação útil, de modo a contribuir para a gestão informacional das demandas por serviços, na administração municipal.

De fato, é atribuição da gestão da informação prover recursos, dentre essas aplicações tecnológicas, de modo a auxiliar as organizações em face dos problemas de informação, cada dia mais complexos e dinâmicos (MARCHIORI, 2002, p. 75-76).

Nessa perspectiva, o aprendizado de máquina e o processamento de linguagem natural têm contribuído com as organizações ao possibilitar a recuperação e extração de informações contidas em documentos, a sumarização e classificação de textos, e a comunicação com clientes e cidadãos, dentre outras atividades.

Até abril de 2021, no PPGGI, havia registro de dois trabalhos abordando o tema “classificação de textos”, concentrando estudos no âmbito da *web* e das redes sociais[2], oportunizando aplicação na administração pública, área ainda não explorada. Esses trabalhos são apresentados no Quadro 1.

QUADRO 1 – PESQUISAS DO PPGGI COM TEMA VINCULADO À CLASSIFICAÇÃO DE TEXTOS

Autor	Título	Descrição
Alcantara (2015)	Recuperação e classificação de informações provenientes da *web* e de redes sociais	O estudo teve por objetivo explorar técnicas de recuperação e classificação de informações na *web* e em redes sociais, a partir da criação de gráficos sociais virtuais para classificação da opinião expressa dos usuários. Esses gráficos, utilizando-se de interações sociais disponíveis nas redes, permitem coletar informações de identificação e da produção dos usuários. As opiniões foram classificadas em positivas, negativas e neutras e os resultados encontrados indicam: uma tendência à opinião negativa; que os usuários parecem trocar informações que vão além das informações ofertadas pelos serviços comerciais de busca e, dessa maneira, esses serviços podem não estar ofertando informações com qualidade relevante aos usuários e; que os gráficos sociais virtuais e a mineração de opinião podem ser utilizados como fatores de classificação complementares ao algoritmo Pagerank, que utiliza a infraestrutura de conexão *web* para classificação da informação.
Nogueira (2015)	Análise de *brand equity* sob a perspectiva do consumidor nas mídias sociais por meio da mineração de opinião e análise de redes sociais	A pesquisa objetivou analisar a equidade das marcas, isto é, a percepção dos consumidores quanto à quatro marcas de cosméticos e quatro marcas da indústria automobilística, a partir de postagens na rede social Twitter. Na preparação dos dados, com vistas a proporcionar análise de cinco dimensões de percepção, foi atribuída polaridade às publicações coletadas. As dimensões analisadas foram conhecimento da marca, lealdade à marca, sentimento percebido, qualidade percebida e associações à marca. A classificação de perfis dos usuários em consumidores e não-consumidores atingiu uma taxa de acerto de 86,5%, com o algoritmo C4.5, e a classificação de polaridade das publicações a precisão de 81,2%, utilizando-se o algoritmo SVM Linear. Os resultados demonstram que a percepção das marcas pode ser analisada a partir de dados das redes sociais, sem a necessidade de aplicação de questionários, possibilitando aos profissionais de *marketing* e à comunidade acadêmica um melhor entendimento sobre como as marcas são percebidas pelos consumidores.

FONTE: A autora (2021)

[2] Disponível em: https://acervodigital.ufpr.br/handle/1884/284, subcomunidades 40001016058P1 e 40001016058P1. Acesso em 30 abr. 2021.

Nos contextos social e econômico, o trabalho justifica-se pelo crescente interesse das organizações na utilização de técnicas de aprendizado de máquina para automatizar atividades e obter vantagem competitiva. Ademais, considerando o crescente aumento de dados desestruturados produzidos no mundo, classificar textos é uma tarefa essencial para recuperação e compreensão da informação.

A escassez de estudos acerca do tema, no âmbito da administração pública, reforça a necessidade de desenvolvimento de pesquisas que contribuam para a modernização e agilização do atendimento ao cidadão no que tange às demandas por serviços públicos. Nas investigações preliminares sobre o contexto para esta pesquisa, não se identificou alguma aplicação de aprendizado de máquina, seja em execução ou em trabalhos acadêmicos, que utilizasse a classificação de demandas textuais em centrais de atendimento ao cidadão.

A Lei Federal n° 13.460/2017, que dispõe sobre a participação, proteção e defesa dos direitos do usuário dos serviços públicos da administração pública, apresenta, como uma das diretrizes para prestação dos serviços públicos, o uso de soluções tecnológicas que simplifiquem processos e procedimentos de atendimento ao cidadão (BRASIL, 2017, p. 2).

Na Prefeitura de Curitiba, as definições investigadas nesta pesquisa, assim como o modelo proposto, poderão auxiliar o desenvolvimento de aplicações voltadas ao aprimoramento e à agilização do atendimento ao cidadão na prefeitura. Estas aplicações incluem: apresentação de uma lista com os serviços mais condizentes à demanda textual que foi registrada; tramitação para o órgão responsável, a partir da classificação efetuada pelo algoritmo e; fornecimento de resposta automática para o cidadão.

Quanto à justificativa pessoal, o desenvolvimento da pesquisa originou-se na percepção da autora, ao realizar a gestão das demandas do 156 do Instituto de Pesquisa e Planejamento Urbano de Curitiba (IPPUC), que parte considerável dos protocolos fora encaminhada ao IPPUC equivocadamente. A autora é analista de sistemas e atua como servidora municipal no IPPUC. Outra motivação para o desenvolvimento da pesquisa é a satisfação da autora em poder ampliar seu conhecimento num tema de grande relevância e que está associado à sua área de formação – Ciência da Computação. Além disso, ao acompanhar e desenvolver atividades vinculadas à Central 156 de Curitiba e constatar o empenho da Prefeitura para aprimorar o atendimento ao cidadão, percebeu uma oportunidade de contribuir

para melhoria do processo de gestão da informação do atendimento ao cidadão, realizado via Central 156.

1.4 DELIMITAÇÃO DA PESQUISA

Atualmente, e conforme exposto, a classificação manual das demandas é realizada em duas etapas, primeiro por assunto e depois por subdivisão, permitindo que o protocolo seja encaminhado ao órgão responsável. Com o objetivo de diminuir a quantidade de assuntos a serem utilizados na classificação automática, esta pesquisa propõe a classificação prévia das demandas por órgão, para que depois possam ser classificadas num dos assuntos do órgão identificado na etapa anterior e, finalmente, numa das subdivisões desse assunto. Assim sendo, a pesquisa limita-se à classificação das demandas por órgão, de modo que possa orientar estudos para classificação das demandas por assunto e subdivisão.

1.5 ESTRUTURA DA DISSERTAÇÃO

A dissertação divide-se em seis capítulos. Após a introdução, que apresenta o tema, a problemática, a questão e os objetivos da pesquisa, assim como a justificativa que ratifica o trabalho, expõe-se a revisão de literatura, no segundo capítulo. A revisão abrange conceitos acerca da gestão da informação na administração pública; do relacionamento governo-cidadão por meio da tecnologia da informação; do aprendizado de máquina; do processamento de linguagem natural; da representação de documentos e da mineração de textos. Nesse capítulo também constam alguns dos trabalhos correlatos à dissertação.

O terceiro capítulo apresenta os encaminhamentos metodológicos, a caracterização e os materiais e métodos da pesquisa. No quarto capítulo é descrito o desenvolvimento do método CRISP-DM para elaboração do modelo proposto, abrangendo compreensão do negócio e dos dados, preparação dos dados, modelagem e avaliação. No quinto são apresentados os resultados e no sexto e último capítulo estão contempladas as considerações finais, contribuições e limitações da pesquisa, bem como as sugestões para trabalhos futuros.

2 REVISÃO DE LITERATURA

Pautando-se em artigos de periódicos, livros, teses e dissertações, este capítulo apresenta os conceitos que fundamentam e promovem melhor compreensão da pesquisa: gestão da informação na administração pública, relacionamento governo-cidadão por meio da tecnologia da informação, aprendizado de máquina, processamento de linguagem natural, representação de documentos e mineração de textos.

2.1 GESTÃO DA INFORMAÇÃO NA ADMINISTRAÇÃO PÚBLICA

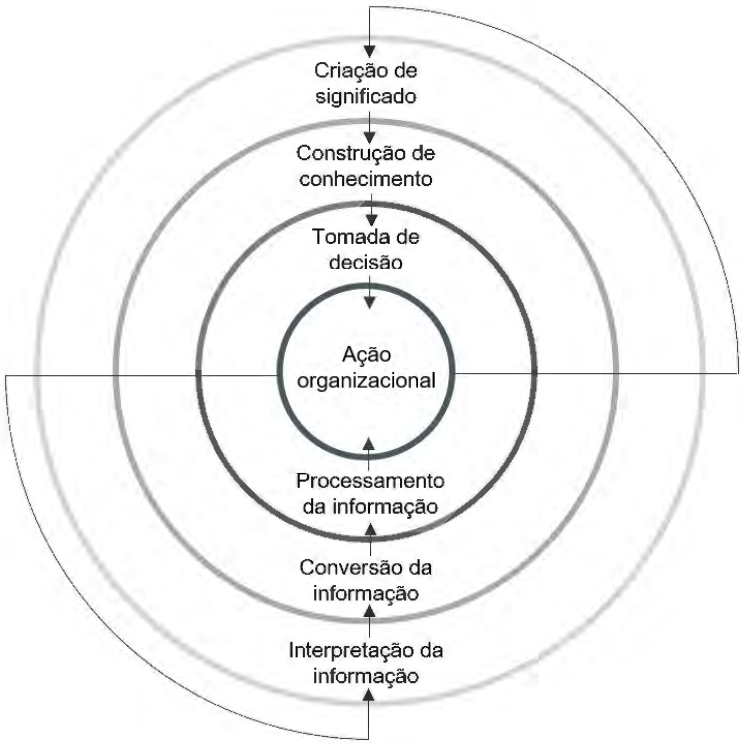
A informação, quando utilizada de modo estratégico, permite que as organizações se adaptem às constantes mudanças do ambiente externo, de modo a melhorar seu desempenho e criar vantagem competitiva (CHOO, 2003, p. 27-28). Também permite que novos conhecimentos sejam gerados e, por conseguinte, o desenvolvimento de novos produtos e serviços, além de possibilitar a tomada de decisão, uma vez que baseia-se em informações associadas aos objetivos da organização (CHOO, 2003, p. 28-29). O autor salienta que o uso estratégico da informação compreende três fases: criação de significado, construção do conhecimento e tomada de decisão (CHOO, 2003, p. 31).

O processo de adaptação às constantes mudanças compreende a fase de criação de significado do uso da informação, na qual notícias e mensagens sobre o ambiente são selecionadas e, a partir da experiência passada, interpretadas sob um ponto de vista consensual entre os membros da organização (CHOO, 2003, p. 28-29).

Na fase de construção do conhecimento, a informação é criada, organizada e processada, de modo a gerar novos conhecimentos, por meio do aprendizado criativo e adaptativo (CHOO, 2003, p. 29). A partir de diálogos e discursos, os membros da organização compartilham seus conhecimentos e a informação é transformada em novos conhecimentos (CHOO, 2003, p. 29). O conhecimento compreende a informação transformada em crenças, conceitos e modelos mentais, a partir do raciocínio e da reflexão (CHOO, 2020; PONJUÁN DANTE, 2004, p. 149), ou seja, é a informação mais valiosa à qual adicionou-se um contexto, um significado, uma interpretação (DAVENPORT; PRUSAK, 1998, p. 19).

Posteriormente, na fase de tomada de decisão, a informação é processada e analisada, levando em consideração vantagens e desvantagens, direcionando a organização a agir de modo oportuno, racional e decisivo (CHOO, 2003, p. 29-30). As fases do uso estratégico da informação, e que compõem o modelo da “organização do conhecimento” definido por Choo (2003, p. 31), estão sintetizadas na Figura 2.

FIGURA 2 – FASES DO USO ESTRATÉGICO DA INFORMAÇÃO: CRIAÇÃO DE SIGNIFICADO, CONSTRUÇÃO DE CONHECIMENTO E TOMADA DE DECISÃO



FONTE: Choo (2003, p. 31)

Para que o êxito seja alcançado nessas três fases, as organizações precisam adotar uma gestão eficiente de informações. Detlor (2010, p. 103) explica que o objetivo da gestão da informação é auxiliar pessoas a acessar, processar e utilizar informações, de modo que possam realizar melhor suas atividades. A esse respeito, Choo (2020) elenca quatro benefícios da gestão da informação: reduzir custos, reduzir incertezas ou riscos, agregar valor aos produtos ou serviços existentes e criar valor a produtos e serviços baseados em informações.

Marchiori (2002, p 74-75), por sua vez, conceitua a gestão da informação sob três enfoques: da administração de empresas, da tecnologia e da ciência da informação.

No enfoque da administração, é um meio para aprimorar os processos com vistas à modernização organizacional, na qual profissionais são capacitados para atuar, além das tradicionais atividades da área, no planejamento e uso estratégico das tecnologias da informação e na qualidade e segurança da informação (MARCHIORI, 2002, p. 74). Na tecnologia, a gestão da informação consiste num recurso potencializado por *hardware, software* e redes de comunicação entre sistemas, geralmente organizacionais (MARCHIORI, 2002, p. 74). E na ciência da informação, a gestão da informação compreende o processo de produção, identificação, coleta, análise, disponibilização, recuperação e uso da informação, objetivando que a informação gerenciada faça “sentido”, isto é, apresente uma finalidade para as pessoas que dela necessitam (MARCHIORI, 2002, p. 75).

Silva e Ribeiro (2009, p. 34) concordam com essa concepção híbrida da gestão da informação, porém, em se tratando do enfoque tecnológico, ressaltam que o ponto central são os sistemas de informação baseados em computador, havendo, portanto, um afastamento da visão tradicional e instrumental da tecnologia. Para o terceiro enfoque, acrescentam que a ciência da informação é constituída por fluxo, organização e comportamento informacional, sendo que esse último contempla a origem, a coleta, a organização, o armazenamento, a recuperação, a interpretação, a transmissão, a transformação e o uso da informação (SILVA; RIBEIRO, 2009, p. 35).

Nesse sentido, a gestão da informação é imprescindível para melhorar a produtividade, a rentabilidade e a competitividade das organizações. Na administração pública, mesmo não havendo competitividade, a abordagem da gestão da informação é análoga, tendo em vista os princípios de legalidade, impessoalidade, moralidade, publicidade e eficiência, enunciados no art. 37 da Constituição da República Federativa do Brasil de 1988 (BRASIL, 2018, p. 25). O princípio da eficiência obsta o desperdício e a má utilização dos recursos públicos, impondo resultados maximizados, com menor custo possível, para satisfação das necessidades coletivas dos cidadãos e dignidade das pessoas com necessidades especiais (JUSTEN FILHO, 2018, p. 108).

Gallo (2010, p. 1) menciona que é tendência na administração pública o volume de dados aumentar além da capacidade de análise. Outrossim, destaca como primordial a definição de políticas públicas que visem à redefinição das necessidades e prioridades de informações, ao planejamento de arquiteturas computacionais

adequadas e flexíveis, à integração de bases de dados e descentralização da criação e à análise de informações.

Esse entendimento é compartilhado por Marchiori (2002, p. 78) ao elencar as habilidades do gestor da informação, dentre estas, auxiliar na definição das necessidades de informação dos usuários, priorizar informações relevantes e de qualidade, utilizar metodologias para desenvolvimento de sistemas de informação e analisar criticamente os recursos de tecnologia quanto ao custo, à qualidade e à complexidade. Gallo (2010, p. 1) evidencia a exigência de conhecimento especializado em gestão da informação, nas três esferas de governo, para implantação de ações programáticas voltadas ao uso estratégico da informação, essencialmente, em benefício dos cidadãos.

Com o objetivo de facilitar a compreensão acerca da gestão da informação, a literatura apresenta modelos que são utilizados pelas organizações para realização desse processo. O modelo proposto por McGee e Prusak (1994, p. 108) é composto por sete etapas: 1 – identificação de necessidades e requisitos de informação; 2 – coleta ou entrada de informação; 3 – classificação e armazenamento de informação; 4 – tratamento e apresentação de informação; 5 – desenvolvimento de produtos e serviços de informação; 6 – distribuição e disseminação de informação e, por fim; 7 – análise e uso da informação.

O modelo de Davenport e Prusak (1998, p. 175-195) é composto por quatro etapas, ou passos: 1 – determinação das exigências da informação; 2 – obtenção de informações, contemplando a exploração do ambiente informacional, classificação, formatação e estruturação da informação; 3 – distribuição e; 4 – uso da informação. Ambos os modelos são utilizados nas explicações que seguem.

Para McGee e Prusak (1994, p. 115) a identificação de necessidades e requisitos é a etapa mais importante do processo de gestão da informação, contudo, muitas vezes, é negligenciada pelos analistas de informação que acreditam saber, de antemão, quais informações são necessárias. Davenport e Prusak (1998, p. 178) afirmam que é preciso acompanhar de perto os gerentes da organização com o propósito de compreender as atividades administrativas e obter conhecimento sobre a informação estruturada e não estruturada, formal e informal, não computadorizada e computadorizada. Ademais, é fundamental o aprofundamento sobre as fontes de informação disponíveis e, ainda, como outras organizações administram suas

informações, de modo a favorecer e enriquecer as entrevistas com os membros da organização (MCGEE; PRUSAK, 1994, p. 116).

A etapa de coleta de informações requer um plano sistemático para aquisição da informação, seja na fonte de origem, seja a partir de cadastramento (MCGEE; PRUSAK, 1994, p. 116-117).

Em geral, as etapas de classificação e armazenamento de informação e tratamento e apresentação de informação ocorrem simultaneamente (MCGEE; PRUSAK, 1994, p. 118). Os autores esclarecem que há necessidade de verificar se o sistema de informações está adaptado à forma de trabalho dos usuários, bem como se a classificação proposta é variada, isto é, permite a recuperação da informação sob vários ângulos. Uma vez que criar classes corretas afeta a maneira como a informação é recuperada, por certo, esse é um processo que exige muita mão-de-obra (DAVENPORT; PRUSAK, 1998, p. 185). A apresentação da informação também requer um grande esforço, pois, quanto mais adequado o formato, em gráficos por exemplo, e mais concisa a informação estiver, mais facilmente será aceita e utilizada (DAVENPORT; PRUSAK, 1998, p. 186-187).

É na etapa de desenvolvimento de produtos e serviços de informação que os usuários podem contribuir, a partir do conhecimento e da experiência que possuem, para a construção de projetos bem-sucedidos (MCGEE; PRUSAK, 1994, p. 119). No entanto, esses indivíduos precisam ser identificados, convidados a participar e ouvidos, sendo suas observações apreciadas pelos analistas de informação. McGee e Prusak (1994, p. 122) classificam esses indivíduos em três categorias: especialistas, com grande conhecimento de sua área de atuação; filtradores, que recebem informações e as filtram conforme qualidade e importância e; fornecedores de redes, que recebem muita informações e utilizam redes informais para distribuí-las.

Quanto à distribuição e disseminação de informação, Davenport e Prusak (1998, p. 190) acreditam que informações individualizadas e contextualizadas, de acordo com a necessidade dos clientes e usuários, despertam maior interesse e satisfação. Para os autores, a distribuição se configura em divulgação aos usuários e busca pelos usuários. Na divulgação, os provedores decidem que tipo de informação deve ser distribuída e a quem, sob o argumento que as pessoas não conhecem o que não sabem. Na busca, as pessoas são estimuladas a procurar pela informação, justificando-se que a informação é distribuída com maior eficiência quando realmente se faz necessária (DAVENPORT; PRUSAK, 1998, p. 190). Em muitas organizações,

a distribuição da informação caracteriza-se como híbrida, com fornecimento de certas informações aos usuários e permissão de acesso a outras.

A última etapa abrange o uso da informação. Ainda que esse processo dependa de como as pessoas procuram, absorvem e compreendem a informação antes de tomar uma decisão, práticas são adotadas pelas organizações em incentivo ao uso da informação (DAVENPORT; PRUSAK, 1998, p. 194-195). Estas práticas incluem a concessão de recompensas e prêmios e a realização de reuniões regulares e de avaliações de desempenho, nas quais são considerados, além dos resultados das decisões dos colaboradores, as informações que levaram a essas decisões (DAVENPORT; PRUSAK, 1998, p. 196-197). Paralelamente, torna-se relevante medir o uso da informação e, em muitas circunstâncias, identificar quem está acessando o quê (DAVENPORT; PRUSAK, 1998, p. 195). Informações com pouco ou nenhum acesso podem ser modificadas ou eliminadas.

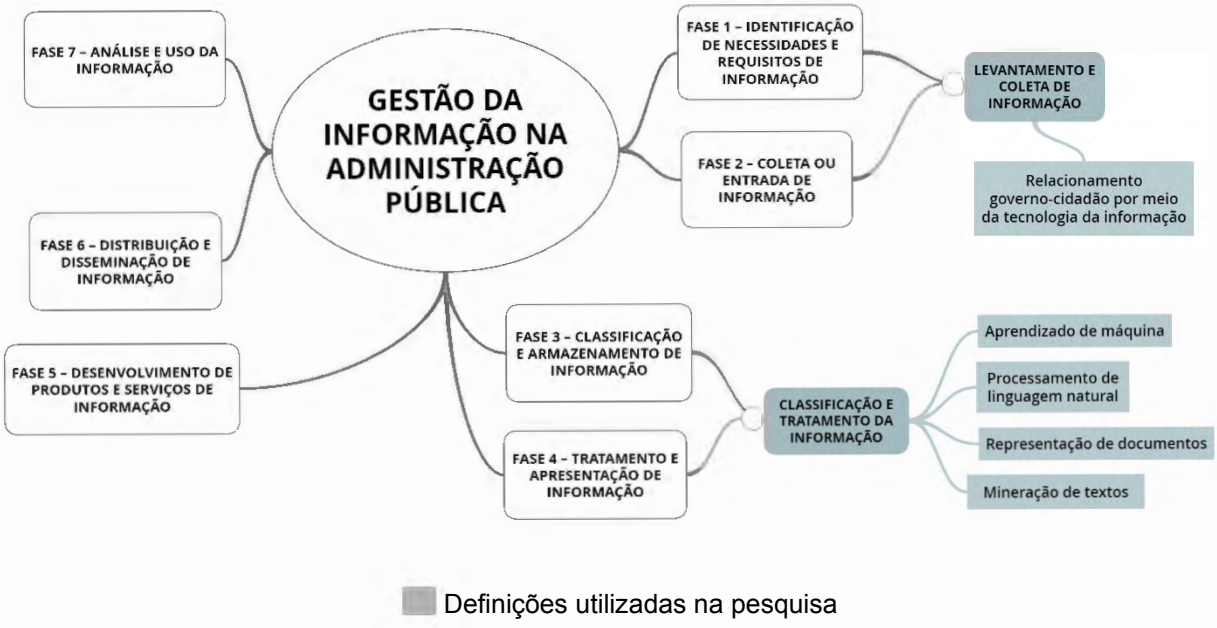
Nesse contexto, os tópicos desta revisão de literatura alinham-se ao modelo de gestão da informação de McGee e Prusak (1994, p. 107-126) da seguinte maneira: as etapas 1 – levantamento de necessidades e 2 – coleta de informações, estão contempladas no tópico que descreve o relacionamento governo-cidadão por meio da tecnologia da informação. É a partir dessa interação que o cidadão fornece o insumo necessário para a gestão das informações, que são as demandas por serviços públicos registradas e classificadas pela Central 156. Essa interação permite, à administração pública, identificação e coleta de informações que o cidadão precisa fornecer em prol do desenvolvimento ou aprimoramento dos sistemas de atendimento.

A etapa 3 – classificação e a etapa 4 – tratamento de informação são consideradas nos tópicos processamento de linguagem natural, representação de documentos, mineração de textos e aprendizado de máquina. Esses tópicos apresentam conceitos que permitem compreender como textos, dados não processados diretamente por computadores, são transformados em novos produtos de informação. Neste estudo, a classificação automática pode ser considerada um novo produto de informação (etapa 5), em auxílio à tomada de decisão dos atendentes no processo de classificação das demandas, bem como para o cidadão, no que tange ao envio automático de determinadas respostas (etapa 6 – distribuição e etapa 7 – uso da informação).

Acrescenta-se que o alinhamento exposto entre a revisão de literatura desta pesquisa e o modelo de gestão da informação de McGee e Prusak (1994, p. 107-126)

seria bastante similar se utilizado o modelo de Davenport e Prusak (1998, p. 175-195). O diagrama da Figura 3 sintetiza as definições investigadas, com marcação daquelas utilizadas na pesquisa.

FIGURA 3 – DIAGRAMA MACRO DAS DEFINIÇÕES INVESTIGADAS E PRODUTOS INFORMACIONAIS ELABORADOS



FONTE: A autora (2020) com base no modelo de McGee e Prusak (1994, p. 108) e auxílio do software GoConqr (EXAMTIME, 2021)

De posse do alinhamento do desenvolvimento da literatura da pesquisa ao modelo de gestão da informação de McGee e Prusak (1994, p. 107-126), exposto na Figura 3, segue para o próximo tópico, o relacionamento governo-cidadão por meio da tecnologia.

2.2 RELACIONAMENTO GOVERNO-CIDADÃO POR MEIO DA TECNOLOGIA DA INFORMAÇÃO

A complexidade das cidades, o avanço das tecnologias de informação e comunicação (TICs) e alterações na legislação, intencionando ampliar a transparência, o controle e a participação da população nas decisões políticas, têm impulsionado mudanças na forma de interação entre governos e cidadãos. Nesse sentido, há um crescente interesse dos governos, organizações da sociedade civil e

cidadãos quanto ao potencial das TICs para aperfeiçoamento das atividades governamentais (RINGOLD *et al.*, 2012, p. 3; WU, 2017, p. 2).

Organizações da sociedade civil e cidadãos podem interagir com o governo de modo a influir na condução da oferta dos serviços públicos, porém, deve haver facilidade de acesso às informações sobre os serviços públicos, demais atos de governo e à legislação, bem como ampliação da compreensão da população acerca dessas informações (RINGOLD *et al.* 2012, p. 6-8). A interação dos cidadãos com o governo segue uma escala crescente de participação, sendo que, nos níveis iniciais, os papéis e as responsabilidades são distintos, geralmente com o governo fornecendo informações, porém, há um intercâmbio de funções nos níveis mais altos, que vai até a deliberação de assuntos por parte da sociedade (DIIRR; ARAUJO; CAPPELLI, 2011, p. 397-398).

Heringer *et al.* (2017, p. 3), com base em Vigoda (2002), consideram que, na teoria, as administrações públicas apresentam cinco níveis de amadurecimento na interação com os cidadãos, que são descritos a seguir.

2.2.1 Níveis de amadurecimento do relacionamento governo-cidadão

O primeiro nível concerne ao governo eletrônico, ou *e-gov,* no qual o cidadão atua como receptor passivo, numa comunicação unidirecional, em que a administração pública é o regulador do processo (HERINGER *et al.*, 2017, p. 6).

As definições de *e-gov* variam, sendo comum associá-lo à utilização da *web* e aos sistemas informatizados para prestação de serviços públicos, de modo a aperfeiçoar o relacionamento entre governo e cidadãos e para prover transparência e prestação de contas, no inglês, *accountability*, das ações governamentais (PRADO *et al.*, 2011, p. 8-9). Consoante a esse entendimento encontra-se o conceito estabelecido por FANG (2002, p. 2) para o *e-gov*

> meio dos governos utilizarem tecnologias inovadoras de informação e comunicação, particularmente baseadas em aplicações web, para fornecer aos cidadãos e empresas acesso mais conveniente a informações e serviços governamentais, com o objetivo de melhorar a qualidade dos serviços e fornecer maiores oportunidades de participação nas regulamentações e nos processos democráticos (FANG, 2002, p. 2, tradução nossa).

Agune e Carlos (2005, p. 1) complementam o raciocínio ao identificar o governo eletrônico como a transição do governo hierarquizado e burocrático para o Estado mais horizontal, inovador e adequado à sociedade do conhecimento. Nesse sentido, o *e-gov* permite que se estabeleça um relacionamento entre governo e cidadãos sem as limitações impostas pelos horários de expediente, de modo a melhorar processos e reduzir gastos, inclusive com o excesso de papel dos documentos (TAVANA; ZANDI; KATEHAKIS, 2013, p. 383).

O governo eletrônico abarca três especificidades de interação: a) divulgação de informações, contemplando impostos, legislação, licenças, registros e obras, dentre outros; b) comunicação, incluindo processos administrativos e diálogo com políticos, autoridades e; c) transações, como a prestação de serviços *on-line* e a publicação de resultados (FANG, 2002, p. 8). Ainda que o modelo burocrático possa prevalecer (AGUNE; CARLOS, 2005, p. 1), o governo eletrônico incorporou-se ao cotidiano de muitas organizações públicas, sendo que os benefícios percebidos pelos cidadãos derivam-se, em especial, dos serviços *on-line* ofertados e, portanto, essa é uma temática que merece avaliação contínua (MA; ZHENG, 2019, p. 2).

Os serviços disponibilizados por meio das TICs são qualificados como *e-serviços* ou, no inglês, *e-services.* Além dos serviços presentes em portais *web*, também fazem parte dos *e-services* aqueles ofertados por outros canais como telefone, tanto fixo como móvel, locais físicos denominados agências, lojas ou centrais atendimento, bem como as centrais de atendimento telefônico (PRADO *et al.,* 2011, p.4). O cidadão escolhe o canal de atendimento conforme sua habilidade de uso das tecnologias, suas perspectivas e experiências anteriores, e a utilização de canais digitais se dá, geralmente, pelo aumento de credibilidade e pelo custo (WU, 2017, p. 5).

Os projetos de *e-gov* implantados com sucesso em vários países trazem como lição que os portais precisam ser abrangentes, permitindo que os cidadãos consigam realizar os processos que necessitam; integrados, de modo que os cidadãos não precisem informar os mesmos dados a cada interação; disponíveis em qualquer equipamento conectado à *web*; fáceis de usar, simplificando e agilizando as transações; acessíveis para pessoas com deficiência; seguros, para proteção e confidencialidade dos dados fornecidos e das transações realizadas e; comunicáveis com outros sites do governo por intermédio de links (FANG, 2002, p. 10).

O estudo realizado por Ma e Zheng (2019, p. 3), com o objetivo de verificar a satisfação dos cidadãos quanto ao desempenho do governo eletrônico em 32 países europeus, concluiu que os sites bem projetados, facilmente navegáveis, amigáveis e abrangentes em cobertura, são os melhores classificados; que, na satisfação geral e específica, a participação é o único indicador com baixa pontuação e; que a satisfação corresponde ao bom desempenho dos governos quanto à oferta de serviços *on-line*, seguida da abertura de participação e da disponibilização de informações.

O estudo de Santos e Rover (2018, p. 222-237) avaliou os portais dos 27 estados brasileiros, apontando, dentre outras considerações, que 65% dos portais não informam e-mails de contato dos governadores e vices; que apenas cinco portais apresentam percentuais de desempenho iguais ou acima de 50%, em se tratando da abertura para participação dos cidadãos quanto à melhoria da prestação dos serviços; que apenas um portal apresenta desempenho igual de 50% no quesito colaboração dos cidadãos envolvendo ferramentas de *e-gov 2.0,* todos os demais ficaram abaixo desse percentual. Nem mesmo para a Lei de Acesso à Informação, mecanismo de transparência passiva, os estados obtiveram bom desempenho, apenas sete apresentaram 100% das variáveis observadas. Quanto aos mecanismos de controle, os portais apresentaram 61,92% das variáveis investigadas. Em média, os portais apresentaram 40,27% das 79 variáveis investigadas, sendo que o portal do estado de São Paulo ficou com a primeira colocação, 61%, o último, do Amapá, com 19% e o do Paraná com 44%.

A pesquisa realizada pela United Nations (2018, p. 27) aponta que o governo eletrônico apresentou crescimento nos últimos anos, entretanto, que a divisão digital persiste, pois 14, dos 16 países com baixos índices de desenvolvimento em governo eletrônico, no inglês, *e-government development index* (EGDI), são africanos. No Gráfico 2 é possível observar a distribuição dos 193 pesquisados para os quatro grupos de classificação do índice.

Os 16 países com baixo EGDI correspondem a 8% e os 71 países com maiores índices, encabeçados por Dinamarca, Austrália e República da Coréia, a 37%. O Brasil ocupa a 44ª posição, com índice alto (UNITED NATIONS, 2018, p. 166).

GRÁFICO 2 – NÚMERO DE PAÍSES AGRUPADOS SEGUNDO ÍNDICE DE DESENVOLVIMENTO DE E-GOV (EGDI) EM 2018

FONTE: United Nations (2018, p. 114)

O segundo nível de amadurecimento refere-se ao governo aberto ou, no inglês, *open government.* Para Heringer *et al.* (2017, p. 3), nesse nível a comunicação passa a ser bidirecional e o cidadão, além de receber informações, interage com o governo como eleitor. Em que pese a relevância dessa relação, o governo aberto intenciona a abertura dos procedimentos administrativos, promovendo a transparência dos dados governamentais e o envolvimento dos cidadãos nos processos administrativos e na tomada de decisão (MEIJER; CURTIN; HILLEBRANDT, 2012, p. 11-12; SCHMIDTHUBER *et al.*, 2017, p. 458).

O termo foi adotado pelo então presidente do Estados Unidos Barak Obama, como indicativo de um novo modelo de administração (NAM, 2011, p. 1; MEIJER; CURTIN; HILLEBRANDT, 2012, p. 1) que incentiva a população a fornecer contribuições ao governo e compartilhar seus conhecimentos com a administração (SCHMIDTHUBER *et al.,* 2017, p. 467). Além da transparência e participação dos cidadãos nas decisões políticas, o governo aberto tem em vista o gerenciamento e o monitoramento da conduta dos servidores públicos, para construção de confiança, responsabilidade, valor público e, consequentemente, uso mais eficiente dos recursos e melhor prestação dos serviços (UNITED NATIONS, 2018, p. 35-39). Para Santos e Rover, 2018, p. 221-222) a busca pela transparência na gestão pública é regra, condição necessária, ainda que não suficiente, para o estabelecimento de um regime democrático, pois também é preciso dispor de mecanismos que possibilitem ao cidadão exercer o efetivo controle social.

O terceiro nível de amadurecimento trata da gestão de relacionamento com o cidadão, no inglês *citizen relationship management* (CiRM), a qual fundamenta-se na

abordagem de gestão de relacionamento com o cliente, no inglês *customer relationship management* (CRM). Nesse nível, o cidadão torna-se cliente e seus dados passam a compor uma fonte de informação e de conhecimento para a administração pública (Heringer *et al.*, 2017, p. 7).

No CRM há um amplo foco no cliente e o objetivo, com auxílio da tecnologia, é dar suporte aos processos de marketing, vendas e serviços, de modo a ampliar o relacionamento com os clientes (SCHELLONG, 2005, p. 2). No CiRM, o foco é o cidadão e a estratégia, habilitada por meio da tecnologia, é aprimorar o relacionamento governo-cidadão, incentivando a cidadania (SCHELLONG, 2005, p. 4). Dessa maneira, o cidadão atua como cliente e o governo como administrador (Heringer *et al.*, 2017, p. 3).

Schellong (2005, p. 4) esclarece que, a partir da experiência no setor privado, para implantação do CiRM no setor público o governo precisa considerar, dentre outras questões: o alto custo envolvido, entre 60 e 130 milhões de dólares; a necessidade de entendimento, em detalhes, dos processos governamentais, ou seja, é preciso alocar recursos humanos nessa atividade; o fato de encontrar novas formas de ofertar os serviços e de criar possibilidades de colaboração e; o desenvolvimento de sistemas de TIC apropriados que consigam absorver as deliberações dos cidadãos.

O quarto e penúltimo nível de amadurecimento é o *e-gov 2.0,* que se baseia na tecnologia *web 2.0* e surge em aprimoramento ao governo eletrônico, no qual o cidadão e governo atuam como parceiros (HERINGER *et al,* 2017, p. 3). Compreende a prestação de serviços governamentais por meio de portais integrados às redes, mídias sociais, *blogs* e outros canais participativos, de modo a ampliar o acesso e fortalecer o relacionamento entre cidadãos, emsas e governo (SUN; KU; SHIH, 2015, p. 504-506). Para os autores, a finalidade maior do *e-gov 2.0* é aproveitar a inteligência coletiva em aprimoramento dos serviços públicos (SUN; KU; SHIH, 2015, p. 511).

Mendes Junior (2018, p. 2) salienta que é necessário agregar mais inteligência aos processos da administração pública, a qual é viabilizada por meio da tecnologia e da participação dos *stakeholders* na gestão e na tomada de decisão das causas públicas. Nesse sentido, o *e-gov 2.0* altera a configuração da prestação dos serviços públicos e do relacionamento entre os *stakeholders* (NAM, 2011, p. 2) ao promover o diálogo mútuo entre cidadãos, governo, empresas e organizações da sociedade civil e ao recuperar e utilizar as demandas apresentadas pela população para tomar

decisões (SUN; KU; SHIH, 2015, p. 507). No entanto, os autores apontam que a transição para o *e-gov 2.0* é complexa, pois requer um planejamento organizacional integrado, não apenas voltado à atualização tecnológica (SUN; KU; SHIH, 2015, p. 504).

Por fim, o último nível *e-democracia,* no inglês *e-democracy*, que apresenta como característica principal o empoderamento do cidadão, em legitimidade às ações governamentais (HERINGER *et al.,* 2017, p. 7). O estudo realizado pelos autores concluiu, entretanto, que há um distanciamento entre a teoria proposta e as efetivas práticas das organizações públicas, essencialmente no que tange à parceria governo-cidadão, transparência, disponibilização de informações relevantes e de interesse da população e utilização do potencial tecnológico.

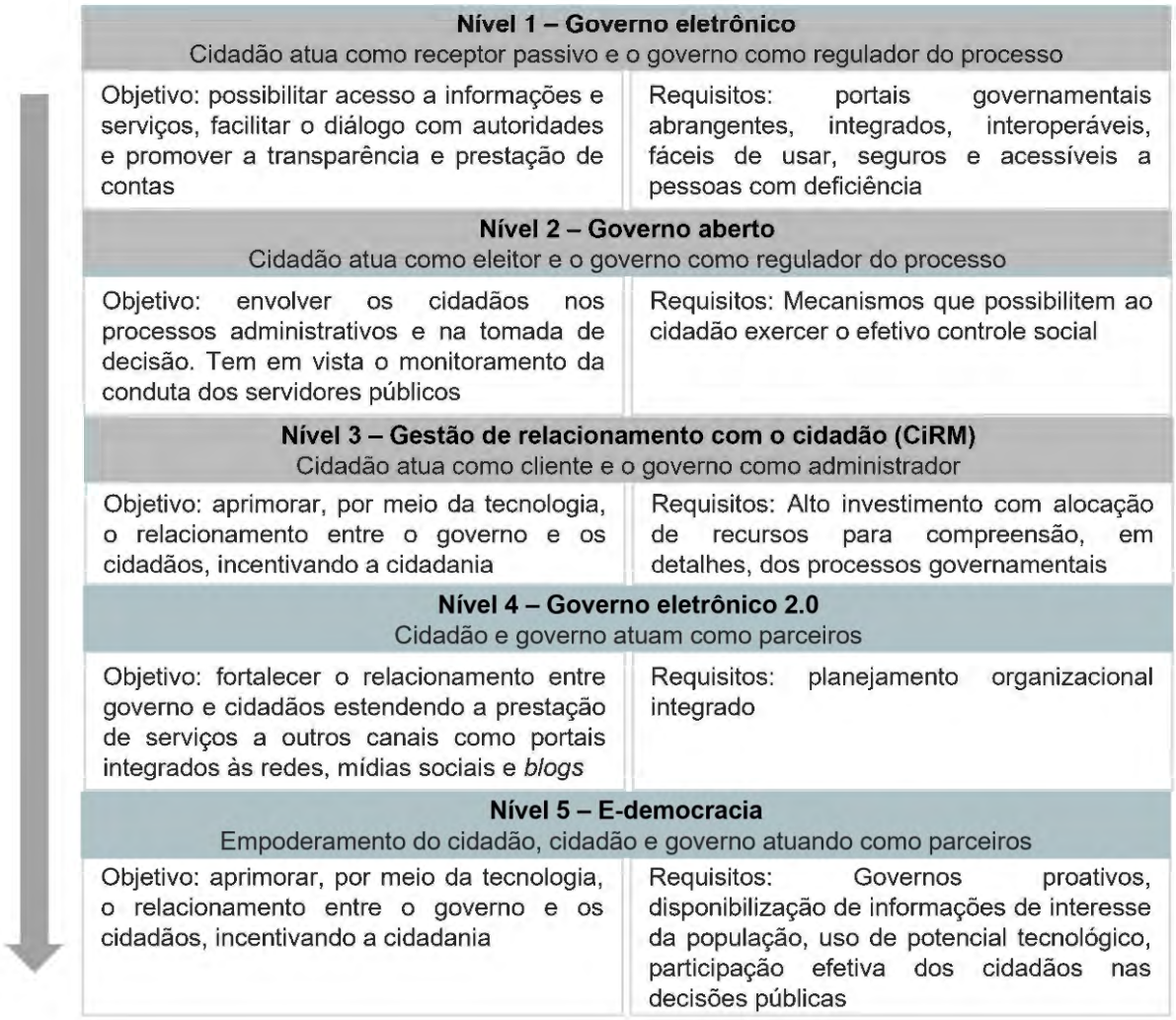
Para Mezzaroba e Bier (2016, p. 210) esse nível é também caracterizado como democracia digital ou em rede e contempla a disponibilização de informações e das contas públicas, votação eletrônica, participação *on-line* dos cidadãos nas questões públicas e interação com políticos. As autoras, em seu estudo, concluem que há ausência de definição acerca da abordagem, assim como Pinho *et al.* (2019, p. 3 e 23), ao explicarem, em suas considerações finais, que a temática não está consolidada e mescla-se aos conceitos de *e-gov* e democracia no geral. Prado *et al.* (2011, p. 5) associam a democracia eletrônica à

> maior participação da sociedade nos processos democráticos e de tomada de decisão do governo por meio de TIC. Relaciona-se também à oferta pelo governo de meios de comunicação, principalmente pela disponibilização de canais de acesso para a diminuição da exclusão digital. Envolve a comunicação entre Estado e cidadão (e a deste com o Estado) e, mais ainda, a comunicação entre todos os participantes do processo político. Trata de *accountability*, e vai além, na implementação da participação ativa do cidadão na elaboração, acompanhamento e controle das políticas públicas, suas práticas e resultados (PRADO et al., 2011, p. 5).

Para que ocorra participação efetiva dos cidadãos nas decisões públicas, governos precisam se tornar facilitadores proativos, fazendo uso das ferramentas de informação e comunicação (SANTOS; ROVER, 2018, p. 222), uma vez que grande parte das soluções limita-se à oferta dos serviços em ambiente *on-line*, porém, desconsiderando um diálogo amplo com a população, entre os próprios cidadãos (DIIRR; ARAUJO; CAPPELLI, 2011, p. 400) e demais stakeholders envolvidos. Santos e Rover (2018, p. 240) avaliam que há falta de preocupação dos governos com ferramentas tecnológicas para relacionamento com os cidadãos e indicam que, no

contexto local, o cidadão ainda é visto como consumidor de um produto final, e não como um parceiro no processo de tomada de decisão. A Figura 4 sintetiza os níveis de amadurecimento do relacionamento entre governo e cidadão abordados neste tópico.

FIGURA 4 – SÍNTESE DOS NÍVEIS DE AMADURECIMENTO DO RELACIONAMENTO ENTRE GOVERNO E CIDADÃO



FONTE: A autora (2021)

Uma vez delimitados os níveis de amadurecimento do relacionamento entre governos e cidadãos, correspondendo às etapas 1 e 2, identificação de necessidades e coleta de informação do modelo de gestão da informação, os próximos tópicos são dedicados às etapas 3 e 4, classificação e tratamento de informação. Esses tópicos contemplam o aprendizado de máquina, o processamento de linguagem natural, a representação de documentos e a mineração de textos.

2.3 APRENDIZADO DE MÁQUINA

O potencial de processamento e armazenamento dos computadores modernos, associado ao *big data*, tem possibilitado a adoção de algoritmos orientados para a aprendizagem computacional, em subsídio à automação e tomada de decisão. Essa prática, dia a dia, torna-se mais presente no contexto organizacional, em virtude da utilização do aprendizado de máquina.

Ainda que computadores não sejam capazes de aprender como os humanos, em certas atividades apresentam desempenho igual ou superior ao de analistas de negócios treinados, com o detalhe que levam muito menos tempo para solucionar problemas (WITTEN *et al.* 2017, p. 26). Ademais, tornou-se impraticável à capacidade humana explorar e analisar dados em tempo real, devido à diversidade, complexidade e velocidade com que são gerados (AWAD; KHANNA, 2015, p. 5).

O aprendizado de máquina compreende o segmento da IA que visa ao desenvolvimento de programas computacionais capazes de aprender, adquirir conhecimento e tomar decisões de forma automática (MONARD; BARANAUSKAS, 2003, p. 39; LOPES; SANTOS; PINHEIRO, 2014, p. 3). Baseia-se em experiências acumuladas e conclusões assertivas (MITCHELL, 1997, p. 2; MONARD; BARANAUSKAS, 2003, p. 39) para questões similares às quais se busca solução. A partir da identificação de padrões e relacionamentos existentes em dados anteriores, o conhecimento adquirido pode ser aplicado a novos elementos, ainda não analisados.

Davenport e Prusak (2000) explicam que o conhecimento se constitui da combinação de experiências, valores e percepções dos especialistas. Sordi (2008, p. 9-12) compartilha desse pensamento ao afirmar que a percepção da informação recebida, associada às aprendizagens anteriores e ao raciocínio, resulta na aquisição do conhecimento.

No contexto organizacional, o conhecimento adquirido passa a incorporar documentos, processos, rotinas e normas (DAVENPORT; PRUSAK, 1998, p. 4). Esse é o conceito de conhecimento tratado pelo aprendizado de máquina e de comum acordo entre os autores Monard e Baranauskas (2003, p. 39), Lopes, Santos e Pinheiro (2014, p. 3) e Witten *et al.* (2017, p. 26), quando afirmam que algoritmos aprendem a partir de padrões, que contemplam experiências ou comportamentos, e aplicam o conhecimento em novos elementos.

As ideias do aprendizado de máquina, descritas nos parágrafos anteriores, são resumidas por Castro e Ferrari (2016, p. 51), ao denominar algoritmos como estruturas.

> A aprendizagem de máquina tem como foco extrair informação a partir de dados de maneira automática. [...] A capacidade de aprender associada às técnicas de aprendizagem de máquina é uma das mais importantes qualidades dessas estruturas. Trata-se da habilidade de adaptar-se ao ambiente de acordo com regras preexistentes, alterando seu desempenho ao longo do tempo. Assim, considera-se aprendizado o processo que adapta o comportamento e conduz a uma melhoria de desempenho de acordo com critérios preestabelecidos (CASTRO; FERRARI, 2016, p. 51).

Witten *et al.* (2017, p. xxiii) relacionam o conceito de aprendizado de máquina ao conceito da mineração de dados. Para os autores, a mineração tem por objetivo extrair informações implícitas, e até então desconhecidas, em conjuntos de dados, cabendo ao aprendizado de máquina fornecer as técnicas básicas, isto é, os algoritmos, para que a mineração possa ser realizada.

Os algoritmos de aprendizado têm origem na união de conhecimentos de várias áreas além da Computação, dentre estas Estatística, Física, Engenharia (LOPES; SANTOS; PINHEIRO, 2014, p. 4), Filosofia e Biologia (CASTRO; FERRARI, 2016, p. 50). A referência como início das atividades de aprendizado de máquina é o trabalho de Arthur Samuel que, entre 1949 e o final da década de 1960, dedicou-se à criação de um programa computacional que aprendesse a jogar Damas, a partir da experiência de jogadores. O programa venceu o quarto melhor colocado dos Estados Unidos à época (MCCARTHY; FEIGENBAUM, 1990, p. 10).

Com o propósito de aprofundar os conceitos relacionados ao aprendizado de máquina e apresentar as abordagens de aprendizagem empregadas nesta pesquisa, torna-se oportuno o detalhamento das formas de inferência existentes, dos paradigmas de aprendizagem mais citados na literatura e dos modos de aprendizado.

2.3.1 Abordagens do aprendizado de máquina

A primeira abordagem envolve a forma de inferência da aprendizagem, que pode ser dedutiva, ou analítica, e indutiva (RUSSELL; NORVIG, 2013, p. 807).

Na forma dedutiva são derivadas, logicamente, novas regras a partir de uma regra geral conhecida (RUSSELL; NORVIG, 2013, p. 807) e as conclusões inferidas

pelo modelo, que se baseia em premissas válidas, são sempre verdadeiras (PRATI, 2006, p. 14).

Na aprendizagem indutiva, em contraposição, regras ou hipóteses são formuladas a partir das descrições e classificações conhecidas de elementos analisados e que auxiliam na classificação de novos elementos (RUSSELL; NORVIG, 2013, p. 887). Portanto, o aprendizado indutivo concebe hipóteses que melhor se adaptam aos elementos observados e as aplica em elementos não avaliados (MITCHELL, 1997, p. 23).

Em outras palavras, a indução utiliza inferência lógica para obter conclusões genéricas a respeito de um conjunto de exemplos, isto é, conceitos ou hipóteses são aprendidos com os exemplos e generalizados para o todo (MONARD; BARANAUSKAS, 2003, p. 40). Os autores acrescentam que, desse modo, as hipóteses geradas podem condizer ou não à verdade. Castro e Ferrari (2016, p. 242) concluem que a generalização é a capacidade do modelo de aprendizado responder corretamente a dados que não fazem parte do conjunto de exemplos.

Ainda que a indução represente o recurso mais utilizado pelos humanos para criar conhecimento novo, no aprendizado de máquina deve ser bem utilizada, com dados de exemplo adequados e suficientes, caso contrário, as hipóteses serão de pouco valor (MONARD; BARANAUSKAS, 2003, p. 40). Nesse aspecto, Awad e Khanna (2015, p. 1) afirmam que os erros, originados na fase de treinamento, devem fomentar a melhoria da generalização do modelo.

Outra abordagem é sobre os paradigmas de aprendizagem. Os mais citados na literatura são: simbólico, conexionista, baseado em exemplos, genético e estatístico.

No paradigma simbólico, os algoritmos aprendem por meio da construção de representações simbólicas de conceitos e da análise de exemplos e contraexemplos (MONARD; BARANAUSKAS, 2003, p. 41). Esse paradigma é utilizado para análise de problemas bem definidos, tendo contribuído principalmente com os sistemas especialistas (LOPES; SANTOS; PINHEIRO, 2014, p. 4). Concentra-se nos processos cognitivos e as representações simbólicas assumem, geralmente, o formato de expressões lógicas e árvores de decisão (MONARD; BARANAUSKAS, 2003, p. 41).

O paradigma conexionista é baseado na hipótese de causa e efeito, na qual “um modelo suficientemente preciso de neurônios biológicos basta para reproduzir a inteligência humana” (LOPES; SANTOS; PINHEIRO, 2014, p. 5). A principal

contribuição desse paradigma são as redes neurais, nas quais a representação é formada por unidades altamente interconectadas (MONARD; BARANAUSKAS, 2003, p. 42). É utilizado para análise de problemas imprecisos que, contudo, podem ser estabelecidos a partir de exemplos (LOPES; SANTOS; PINHEIRO, 2014, p. 5).

No paradigma baseado em instâncias ou exemplos, no inglês *instance based*, para classificar um exemplo, são utilizados exemplos similares e com classe conhecida (MONARD; BARANAUSKAS, 2003, p. 42). Outrossim, esse tipo de aprendizado é *lazy*, ou preguiçoso, porque precisa manter os exemplos similares na memória para classificar novos exemplos (MONARD; BARANAUSKAS, 2003, p. 42). A contribuição desse paradigma é o algoritmo k-vizinhos mais próximos, no inglês k-*nearest neighbors* (k-NN) (CASTRO; FERRARI, 2016, p. 51-52; GOLDSCHMIDT; PASSOS, 2005, p. 98; MONARD; BARANAUSKAS, 2003, p. 42).

No paradigma genético existe uma competição entre os elementos de classificação para realizar a predição, e aqueles que apresentam desempenho ruim são descartados, como na teoria de Darwin, onde sobrevivem os que mais se adaptam ao ambiente (MONARD; BARANAUSKAS, 2003, p. 43).

No paradigma estatístico, o objetivo é a obtenção de um conceito a ser induzido, que apresente bom desempenho na classificação (MONARD; BARANAUSKAS, 2003, p. 41) e, para tanto, busca-se pelos melhores parâmetros para ajuste do modelo (PRATI, 2006, p. 19). A principal contribuição desse paradigma são os algoritmos bayesianos (MONARD; BARANAUSKAS, 2003, p. 41), porém, alguns autores também incluem as redes neurais, uma vez que no treinamento é necessário encontrar valores apropriados para os pesos da rede (PRATI, 2006, p. 19).

A última abordagem é quanto ao modo de aprendizado, sendo os aprendizados não supervisionado, semissupervisionado, supervisionado e por reforço, os mais citados dentre os autores pesquisados.

O aprendizado não supervisionado compreende algoritmos projetados para descoberta de informações implícitas em grandes conjuntos de dados (CASTRO; FERRARI, 2016, p. 40), nos quais as saídas, ou *feedbacks*, são desconhecidos, isto é, não fornecidos ao algoritmo (AWAD; KHANNA, 2015), p. 7; RUSSELL; NORVIG, 2013, p. 808; GOLDSCHMIDT; PASSOS, 2005, p. 55). As tarefas mais comuns de aprendizagem não supervisionada são a associação e o agrupamento.

Na associação procura-se descobrir padrões que denotem relações entre as descrições, ou atributos, dos elementos analisados (FELDMAN; SANGER, 2007, p.

25). Essa técnica extrai relações entre atributos de um conjunto de dados, no qual “se A então B”, indicando que quando o elemento A ocorre, o B tende a ocorrer (VASQUES *et al.,* 2017, p. 12). Para Goldschmidt e Passos (2005, p. 59) a associação consiste em “encontrar conjuntos de itens que ocorram simultaneamente e de forma frequente em um banco de dados”. Os autores trazem o exemplo clássico da rede de supermercados que descobriu a compra associada de fraldas e cervejas, antes dos finais de semana que havia transmissão de jogos (GOLDSCHMIDT; PASSOS, 2005, p. 59).

No agrupamento, ou *clustering*, o conjunto de elementos é particionado em grupos, conforme similaridades encontradas pelo algoritmo sem, contudo, que houvesse classificação prévia desses elementos (CASTRO; FERRARI, 2016, p. 41-42). Os grupos identificados pelo algoritmo passam a representar as classes do conjunto de elementos.

O aprendizado semissupervisionado inclui algoritmos para classificação em situações onde existem poucos elementos classificados previamente e muitos dados inconsistentes (RUSSELL; NORVIG, 2013, p. 808). Considerando-se que a classificação de dados exige mão de obra humana qualificada, atividade de alto custo, e que dados não classificados são de fácil obtenção e baixo custo, o aprendizado semissupervisionado é uma opção viável (AWAD; KHANNA, 2015, p. 8). Desse modo, é utilizada uma combinação entre um pequeno número de elementos classificados e um grande conjunto de dados sem classificação, que tenta gerar uma hipótese de classificação (AWAD; KHANNA, 2015, p. 8).

No aprendizado por reforço, o algoritmo aprende com base em recompensas ou punições (RUSSELL; NORVIG, 2013, p. 808). Por exemplo, ao executar uma tarefa adequadamente recebe pontos condizentes a uma recompensa, caso contrário deixa de receber os pontos, cabendo ao algoritmo decidir qual das ações de reforço são responsáveis pelos acertos (AWAD; KHANNA, 2015, p. 8).

No aprendizado supervisionado, um conjunto de elementos classificados, também denominados exemplos, tuplas, casos, dados de treinamento ou instâncias, auxiliam na classificação de novos elementos.

2.3.2 Aprendizado supervisionado

Esse aprendizado recebe a denominação de supervisionado visto que um supervisor humano define as classes dos exemplos de treinamento, de modo a orientar o processo de aprendizagem (MANNING; RAGHAVAN; SCHÜTZE, 2009, p. 256).

O aprendizado supervisionado envolve a concepção de um modelo de conhecimento, formulado pelo algoritmo a partir de exemplos, dispostos em pares ordenados, que apresentam uma entrada e a saída desejada (GOLDSCHMIDT; PASSOS, 2005, p. 55; RUSSELL; NORVIG, 2013, p. 808). Desse modo, o algoritmo analisa os exemplos e cria um mapeamento das entradas para as saídas (RUSSELL; NORVIG, 2013, p. 808) e que é utilizado para produção das saídas, ao receber valores especificados em novas entradas (GOLDSCHMIDT; PASSOS, 2005, p. 55).

Nessa abordagem, as saídas desejadas constituem o atributo diferencial, também denominado rótulo ou classe, que caracteriza cada exemplo e contempla “a meta que se deseja aprender e poder fazer previsões a respeito” (MONARD; BARANAUSKAS, 2003, p. 44). Seguindo esse entendimento, Awad e Khanna, 2015, p. 6) definem a aprendizagem supervisionada como

> um mecanismo que infere a relação implícita entre os dados observados (também chamados de dados de entrada) e uma variável de destino (uma variável dependente ou rótulo) suscetível à previsão [...]. A aprendizagem utiliza dados de treinamento rotulados (exemplos de treinamento) para sintetizar um modelo que tenta generalizar a relação implícita entre os vetores de características (entrada) e os indicadores de supervisão (saída) (AWAD; KHANNA, 2015, p. 6, tradução nossa).

Ao utilizar os dados de treinamento, o propósito do algoritmo, também denominado indutor, é a construção de um classificador que determine a classe de exemplos que não contenham o rótulo, de forma correta (MONARD; BARANAUSKAS, 2003, p. 40). Para rótulos de classes discretas, ou nominais, essa tarefa é conhecida por classificação e, para valores contínuos, por regressão.

Castro e Ferrari (2016, p. 241), numa acepção análoga, consideram que quando cada exemplo do conjunto de dados possui uma classe associada, retratando o histórico dos registros, o objetivo do algoritmo é a construção de um modelo que consiga predizer a saída para elementos não rotulados. Esses autores denominam a tarefa de classificação que envolve valores contínuos de estimação.

O modelo compreende uma estrutura que resume o conjunto de dados, sendo utilizado para predizer as saídas, e que pode ser ajustado parametricamente, isto é, por meio de um conjunto finito de parâmetros, ou não parametricamente, com um conjunto infinito de parâmetros (AWAD; KHANNA, 2015, p. 4). Para os autores, modelos não paramétricos são simples, contudo, necessitam de conjuntos de dados maiores para obter conclusões precisas. Para Barba (2020) esse modelo consiste numa representação matemática que acumula e se modifica, à medida que mais aprende com os dados de treinamento, no sentido de melhorar continuamente para resolver uma atividade preestabelecida.

Nesse contexto, a regressão é utilizada quando a classe, isto é, resultado esperado, é numérico (WITTEN *et al.,* 2017, p. 128; CASTRO; FERRARI, 2016, p. 325; GOLDSCHMIDT; PASSOS, 2005, p. 13) e os atributos dos exemplos de treinamento também o são (WITTEN *et al.,* 2017, p. 128). O objetivo é a obtenção de uma função que mapeie a relação entre a(s) variável(is) de saída, também denominadas dependentes, preditas, explicadas ou de resposta, e as preditoras, também chamadas de variáveis de controle, independentes ou de entrada (CASTRO; FERRARI, 2016, p. 325).

A classificação compreende a atribuição de rótulos categóricos predefinidos, obtidos a partir de exemplos de treinamento, a novos registros, de modo a prever a classe desses registros (GOLDSCHMIDT; PASSOS, 2005, p. 13; CASTRO; FERRARI, 2016, p. 158). Nessa tarefa, o algoritmo recebe os exemplos de treinamento rotulados e cria um modelo para classificação correta de novos registros (CASTRO; FERRARI, 2016, p. 158).

Usualmente, um conjunto de exemplos é dividido em dois subconjuntos: o de treinamento, usado para o aprendizado do modelo e o de testes, para mensurar a precisão do modelo (MONARD; BARANAUSKAS, 2003, p. 44; RUSSELL; NORVIG, 2013, p. 809). Consequentemente, para que a avaliação seja isenta, os exemplos utilizados na construção do modelo não devem fazer parte do conjunto de testes (GOLDSCHMIDT; PASSOS, 2005, p. 50). Em outras palavras, o conjunto de treinamento é usado para construir um classificador que, então, é utilizado para prever a classificação dos exemplos do conjunto de testes (BRAMER, 2016, p. 80).

Dentre as técnicas de treinamento e teste comumente utilizadas estão *holdout* e validação cruzada. Com *holdout*, o conjunto de exemplos é dividido em duas partes mutuamente exclusivas, utilizando-se um percentual fixo de exemplos para

treinamento e o remanescente para teste. A parte reservada para o treinamento é geralmente maior, sendo comum empregar 2/3 para treinamento e 1/3 para teste (MONARD; BARANAUSKAS, 2003, p. 53). Na validação cruzada, ou *cross-validation* no inglês, os exemplos são aleatoriamente divididos em *n* partes, ou *folds*, mutuamente exclusivas e de tamanhos aproximadamente iguais. São utilizados para treinamento os exemplos das (*n* – 1) partes, sendo o teste realizado na parte remanescente. Esse processo é repetido *n* vezes, considerando-se cada vez uma parte diferente para teste. O erro corresponde à média dos erros calculados em cada uma das *n* repetições do processo (MONARD; BARANAUSKAS, 2003, p. 53).

Os algoritmos de classificação são adequados para problemas com limites bem definidos, isto é, quando as entradas formam um conjunto de atributos específicos e a saída é uma variável categórica alvo (AWAD; KHANNA, 2015, p. 21; LAROSE; LAROSE, 2014, p. 10). O algoritmo examina os exemplos de entrada, cada qual contendo informações sobre a variável alvo (LAROSE; LAROSE, 2014, p. 10) e identifica padrões. Num segundo momento, combina esses padrões às novas entradas e, se houver correspondência, um padrão preditivo é associado à entrada (AWAD; KHANNA, 2015, p. 21).

O algoritmo, também denominado programa de aprendizado ou simplesmente indutor, consiste na extração, a partir de exemplos categorizados, de um bom classificador, também chamado de hipótese ou descrição de conceito (MONARD; BARANAUSKAS, 2003, p. 43), que seja capaz de generalizar além do conjunto de dados de treinamento (AWAD; KHANNA, 2015, p. 3). Desse modo, no conceito de Awad e Khanna (2015, p. 2), um classificador é

> um método que recebe uma nova entrada, como uma instância não rotulada de uma observação ou característica, e identifica uma categoria ou classe à qual pertence. Muitos classificadores comumente usados empregam inferência estatística (medida de probabilidade) para categorizar o melhor rótulo para uma determinada instância (AWAD; KHANNA, 2015, p. 2, tradução nossa).

Nesse entendimento, um exemplo é um par ordenado $(x_i, f(x_i))$ onde x_i é a entrada e $f(x_i)$ é a saída, sendo o objetivo do algoritmo de indução gerar uma função h que melhor aproxime f, via de regra desconhecida e, desse modo, h é a hipótese sobre a função f, ou seja, $(x_i) \approx f(x_i))$ (MONARD; BARANAUSKAS, 2003, p. 45).

Outro ponto importante a ser analisado na tarefa de classificação diz respeito à prevalência de classes num conjunto de exemplos. Em conjuntos de dados com classes de mesma relevância, a distribuição dos exemplos nas classes deve ser realizada de forma igualitária, uma vez que classes mais frequentes, ou majoritárias, podem dominar na predição dos elementos não rotulados (CASTRO; FERRARI, 2016, p. 272). No aprendizado, além da prevalência de classes, é também utilizado o termo desbalanceamento de classes (MONARD; BARANAUSKAS, 2003, p. 44).

Dentre as abordagens propostas para contornar o desbalanceamento estão dois métodos simples de reamostragem, o *undersampling* e o *oversampling*. *Undersampling* consiste na eliminação de exemplos das classes majoritárias e *oversampling* na inclusão de exemplos para as classes minoritárias, que é realizada por meio de replicação de exemplos (PRATI, 2006, p. 89).

Ainda que esses métodos possam apresentar como limitações a eliminação de exemplos potencialmente importantes, quando utilizado *undersampling*, e o superajustamento do modelo classificador para *oversampling* (PRATI, 2006, p. 89; LIU 2004, p. 17), ambos são utilizados para melhorar a precisão da classificação de conjuntos de dados com classes desbalanceadas (HAN; KARYPIS; KUMAR, 1999, p. 383). Esses métodos possibilitam que exemplos sejam eliminados ou acrescidos aleatoriamente, até que a distribuição entre as classes seja igual, fazendo com que as classes minoritárias tornem-se representativas (HAN; KARYPIS; KUMAR, 1999, p. 383-384).

No estudo realizado por Prati (2006) foram realizados experimentos em conjuntos de dados artificiais e naturais, dispostos nas classes majoritária e minoritária, tendo como uma das variáveis de controle a proporção de exemplos em cada classe. As proporções, para os diversos experimentos, foram alteradas por meio dos métodos baseados em *under* e *oversampling*, de modo a permitir análise do desempenho dos algoritmos de aprendizado. O autor concluiu que *oversampling* apresenta melhores resultados quando em se tratando de classes desbalanceadas e sobreposição de exemplos entre as classes *oversampling* e ainda que esse método é bem competitivo quando comparado a outros métodos de *oversampling* mais sofisticados.

No estudo de Liu (2004), como no de Prati (2006), foram definidas duas classes, a minoritária, com menor número de exemplos, e a majoritária, formada pelo agrupamento das demais. Os métodos de reamostragem foram aplicados e diferentes

experimentos de classificação realizados com os algoritmos Naïve Bayes, k-NN e SVM. O autor concluiu que a reamostragem pode melhorar significativamente o desempenho dos algoritmos.

Para que o conjunto de exemplos possa ser submetido ao algoritmo indutor, normalmente é utilizado o formato atributo-valor, isto é, uma tabela na qual cada exemplo (E) com a respectiva classe (C), ocupa uma linha, cada atributo (A), uma coluna (MONARD; BARANAUSKAS, 2003, p. 44; AWAD; KHANNA, 2015, p. 3) e os valores dos atributos ocupam as células da tabela.

Uma vez delimitado o conceito de aprendizado de máquina e das abordagens que o permeiam, apresenta-se a classificação de textos e os algoritmos de aprendizagem mais utilizados nessa tarefa.

2.3.3 Classificação de textos

Classificar coisas é algo inerente aos seres humanos. O cérebro classifica, ou categoriza, as estruturas que espelham o ambiente externo (LIMA, 2010, p. 110) e, mesmo que a interação com o ambiente ocorra a nível de objetos individuais, no raciocínio, ocorre a nível de categorias (RUSSELL; NORVIG, 2013, p. 520). A classificação organiza e simplifica a base do conhecimento por herança, visto que as subclasses herdam propriedades das classes, e assim por diante, definindo uma taxonomia ou hierarquia taxonômica (RUSSELL; NORVIG, 2013, p. 520).

No aprendizado de máquina, a tarefa de classificação é essencial para replicação da tomada de decisão humana, que ocorre por meio de análises preditivas sobre dados (AWAD; KHANNA, 2015, p. 21), incluindo aqueles em formato textual. Utilizando, exclusivamente, o conteúdo dos próprios arquivos como insumo, a classificação de textos apresenta-se como um tópico importante no aprendizado de máquina atualmente (WITTEN *et al.*, 2017, p. 6).

Estudos acerca da categorização automática de textos iniciaram na década de 1960, utilizando-se vocabulário controlado para indexar a literatura científica (FELDMAN; SANGER, 2007, p. 64). Até o final dos anos 1980, a abordagem mais popular era a engenharia do conhecimento (SEBASTIANI, 2002, p. 2; WITTEN, 2004, p. 6), contudo, na década de 1990, a classificação automática de textos predomina e se consolida (FELDMAN; SANGER, 2007, p. 64).

A classificação de textos é também conhecida por categorização de textos (FELDMAN; SANGER, 2007, p. 64; RUSSELL; NORVIG, 2013, p. 996) e classificação estatística de texto, se o método de aprendizagem for estatístico (MANNING; RAGHAVAN; SCHÜTZE, 2009, p. 255). Corresponde ao processo de classificar corretamente a categoria (ou categorias) de cada exemplo, dado um conjunto de categorias (assuntos ou tópicos) e uma coleção de documentos de texto (FELDMAN; SANGER, 2007, p. 64; HAN; KARYPIS; KUMAR, 1999, p. 383).

Na classificação de textos tem-se como conceitos fundamentais que cada texto individual é denominado documento, ou exemplo, e que as palavras ou termos que o compõem compreendem seus atributos ou características, de modo a formar uma amostra de dados a ser submetida às tarefas de aprendizado (WEISS; INDURKHYA; ZHANG, 2015, p. 4). Processando quantidades de dados que os humanos sequer imaginam conseguir organizar ou associar, os algoritmos identificam padrões em combinações de palavras recorrentes, que permitem predizer classes (WEISS; INDURKHYA; ZHANG, 2015, p. 4).

As variáveis utilizadas para nomear ou atribuir os rótulos, na maioria das vezes, são nominais ou categóricas, isto é, variáveis alfanuméricas que geralmente apresentam um conjunto finito de situações possíveis como, por exemplo, o estado civil, contemplando os atributos solteiro, casado, viúvo e divorciado (GOLDSCHMIDT; PASSOS, 2005, p. 24-25).

Existem dois enfoques principais da categorização de textos. O primeiro concerne à engenharia do conhecimento, na qual o conhecimento do especialista, acerca das categorias de determinado domínio, é codificado no sistema sob a forma de regras de classificação (FELDMAN; SANGER, 2007, p. 64; SEBASTIANI, 2002, p. 2), sendo associado ao paradigma simbólico de aprendizagem. O segundo enfoque refere-se ao aprendizado de máquina, no qual um processo indutivo possibilita a criação de um classificador automático, que aprende a partir de exemplos com categorias predefinidas (FELDMAN; SANGER, 2007, p. 64).

Para os autores Feldman e Sanger (2007, p. 64), no domínio da gestão de documentos, o enfoque da engenharia do conhecimento supera o enfoque do aprendizado de máquina, apesar de exigir um quadro enorme de recursos humanos qualificado, ou seja, pessoas com conhecimento especializado para criar e manter as regras de codificação.

Sebastiani (2002, p. 2), ao contrário, defende que a classificação, por meio do aprendizado de máquina, apresenta precisão comparável à alcançada por especialistas humanos e uma economia considerável com força de trabalho, uma vez que não carece da intervenção de especialistas de domínio. Manning, Raghavan e Schütze (2009, p. 255), em conformidade a esse raciocínio, defendem que rotular um conjunto de documentos é mais fácil do que criar regras.

Outra vantagem considerável do aprendizado automático, em comparação às regras, é que os algoritmos disponíveis podem ser aplicados a diferentes domínios e reaplicados quando há alteração nas categorias de determinado domínio, para o qual o classificador foi construído (SEBASTIANI, 2002, p. 8).

O conjunto de classes de determinado domínio pode assumir a seguinte tipificação: binária, ou *binary* no inglês, rótulo único, ou *single-label*, e múltiplos rótulos ou *multilabel*.

A classificação binária é a mais simples e mais comum, na qual existem apenas dois rótulos possíveis a serem atribuídos ao novo documento e cada documento deve pertencer a uma das duas classes (MANNING; RAGHAVAN; SCHÜTZE, 2009, p. 306). Nesse caso, geralmente, existe a categoria e o seu negativo (FELDMAN; SANGER, 2007, p. 67; SEBASTIANI, 2002, p. 3), isto é, o documento pertence ou não à categoria, como no caso de *spam*, no qual o e-mail é ou não *spam*.

Na classificação de rótulo único há mais de duas categorias e cada documento pertence a apenas uma delas. Para Feldman e Sanger (2007, p. 67), essa classificação é, frequentemente, uma generalização da binária, na qual as classes são mutuamente exclusivas (MANNING; RAGHAVAN; SCHÜTZE, 2009, p. 306), sendo também conhecida como categorias sobrepostas, ou no inglês *overlapping categories* (SEBASTIANI, 2002, p. 3).

A classificação de múltiplos rótulos compreende a sobreposição de categorias para um mesmo documento, isto é, um único documento pode pertencer a várias categorias simultaneamente (FELDMAN; SANGER, 2007, p. 67; MANNING; RAGHAVAN; SCHÜTZE, 2009, p. 306). Para Sebastiani (2002, p. 2), no caso de múltiplos rótulos, o número de categorias que um documento pode assumir vai de 0 ao número máximo de categorias disponíveis.

Feldman e Sanger (2007, p. 70) entendem que o processo de classificação de textos abarca quatro questões principais. Primeiro é preciso decidir as categorias que serão utilizadas, depois é necessário selecionar um conjunto de treinamento para

cada uma das categorias. Os autores apontam que, como regra geral, cerca de 30 exemplos são necessários para cada categoria. Na sequência, está a decisão sobre quais atributos representam significativamente cada um dos exemplos. Finalmente, é necessário decidir qual ou quais algoritmos serão utilizados.

2.3.4 Algoritmos de classificação de textos

Dentre os algoritmos utilizados na tarefa de classificação de textos destacam-se, pela simplicidade e ampla utilização, os baseados em árvores de decisão, k-vizinhos mais próximos (k-NN) e no Teorema de Bayes.

Uma árvore de decisão, ou *decision tree* no inglês, cria uma estrutura no formato de árvore, iniciando pelo nó raiz, ao qual é aplicado um teste de um atributo, seus possíveis resultados formam ramos e os nós folhas representam as classes. Os caminhos entre o nó raiz e cada nó folha definem as regras de classificação (CASTRO; FERRARI, 2016, p. 275-276).

Aggarwal e Zhai (2012, p. 176), de maneira análoga, argumentam que uma árvore de decisão decompõe hierarquicamente o espaço de exemplos em condições, nas quais são utilizados atributos que marcam as divisões do espaço. No contexto da classificação de documentos, as condições representam a presença ou ausência de palavras ou termos no documento.

Para identificar a classe à qual o exemplo não rotulado pertence é necessário percorrer os caminhos da árvore, de cima para baixo, conforme os atributos do exemplo, até atingir um nó folha, que corresponde à classe predita para o registro em questão (AGGARWAL; ZHAI, 2012, p. 176; CASTRO; FERRARI, 2016, p. 276).

Russell e Norvig (2013, p. 821) consideram que algoritmos de árvores de decisão, geralmente, são os primeiros a serem experimentados para a tarefa de classificação, inclusive, vários pacotes de software baseados nesse método foram elaborados visando atender demandas industriais e comerciais.

Uma propriedade importante das árvores de decisão é que são fáceis de compreender, isto é, um ser humano consegue entender a razão da saída do algoritmo de aprendizagem (CASTRO; FERRARI, 2016, p. 276; RUSSELL; NORVIG, 2013, p. 821) sendo, por essa razão, denominados algoritmos caixa branca (CASTRO; FERRARI, 2016, p. 276), o que não acontece com as redes neurais (RUSSELL; NORVIG, 2013, p. 821).

Um dos algoritmos de árvore de decisão utilizado é o J48. Esse algoritmo consiste na implementação do algoritmo C4.5, proposto por Quinlan em 1993 (WITTEN *et al.,* 2017, p. 158). No C4.5 a árvore inicia com um único nó representando os dados de treinamento. Se todos os objetos pertencem à mesma classe, o nó torna-se uma folha e é rotulado com aquela classe, caso contrário, é calculado o atributo que melhor separa as classes em classes individuais, e que se torna o atributo teste ou atributo de decisão do nó. Para cada valor conhecido do atributo teste, os dados são particionados seguindo esses valores. O processo é realizado recursivamente formando uma árvore de decisão e, uma vez que um atributo apareceu em um nó, não precisa mais ser considerado nos seus descendentes. O particionamento recursivo é interrompido quando todos os objetos de um nó pertencem à mesma classe ou quando não há mais atributos para os quais os objetos precisem ser particionados (CASTRO; FERRARI, 2016, p. 283).

Algoritmos baseados em instâncias, como o k-NN, ao classificar um novo registro, calculam a distância entre esse registro e cada um dos exemplos existentes na base de referência, identificando aqueles k registros mais próximos ou similares (AWAD; KHANNA, 2015, p. 14; GOLDSCHMIDT; PASSOS, 2005, p. 99; LAROSE; LAROSE, 2014, p. 152). Dessa maneira, a classe da maioria dos k vizinhos é atribuída ao novo exemplo e, no caso de empate, uma classe é atribuída aleatoriamente (CASTRO; FERRARI, 2016, p. 272).

A decisão baseada na classe da maioria dos k vizinhos mais próximos é denominada voto majoritário (CASTRO; FERRARI, 2016, p. 168; WANG, 2006, p. 2), na qual assume-se que os k vizinhos possuem o mesmo peso na decisão da classe, independentemente de suas distâncias ao registro não rotulado (WANG, 2006, p. 2). Outra possibilidade é atribuir diferentes pesos aos k vizinhos, baseando-se nas distâncias em relação ao registro não rotulado. Nesse caso, os k vizinhos mais próximos recebem pesos maiores e a classe que apresentar maior soma, entre os k vizinhos, é atribuída ao registro não rotulado (WANG, 2006, p. 2).

Para identificar os exemplos similares, geralmente, é empregada a função de distância euclidiana (LAROSE; LAROSE, 2014, p. 153) ou a função do cálculo do cosseno (ARANHA, 2007, p. 24). Como a atribuição se dá em virtude da quantidade de exemplos similares, é importante que exista equilíbrio da quantidade de exemplos de cada classe no conjunto de treinamento, pois, caso contrário, as classes mais frequentes podem dominar a predição (CASTRO; FERRARI, 2016, p. 272).

O k-NN é um método preguiçoso de aprendizagem, *lazy* no inglês, pois adia o processamento dos dados de treinamento até que uma consulta seja realizada, isto é, até que um registro não rotulado precise de classificação (WANG, 2006, p. 2). Para Castro e Ferrari (2016, p. 272), o aprendizado é *lazy* por não haver treinamento *a priori* do modelo, uma vez que a saída é calculada somente no momento de atribuição da classe ao novo objeto e, consequentemente, não há criação ou ajuste prévio do modelo (CASTRO; FERRARI, 2016, p. 51-52).

Esse processo, normalmente, envolve o armazenamento de todo o conjunto de treinamento na memória (MITCHELL, 1997, p. 230; WANG, 2006, p. 2), tornando o algoritmo desvantajoso (MITCHELL, 1997, p. 245). Ainda assim, o k-NN é amplamente utilizado e apresenta bons resultados em várias situações (CASTRO; FERRARI, 2016, p. 168). Para Wang (2006, p. 2) o k-NN se mostra um método efetivo de classificação, uma vez que tem sido utilizado com sucesso em muitas aplicações do mundo real.

Na classificação de textos, Weiss, Indurkhya e Zhang (2015, p. 45) consideram o k-NN um dos métodos mais relevantes, embora necessite de grande potencial computacional (WEISS; INDURKHYA; ZHANG, 2015, p. 53). Sebastiani (2002, p. 29) também afirma que o k-NN é bastante efetivo, no entanto, avalia o tempo de processamento como uma desvantagem e Manning, Raghavan e Schütze (2009, p. p. 290) consideram o k-NN menos eficiente que outros métodos.

Um dos algoritmos baseado em instâncias utilizado é o IBk (*Instance Based k-Nearest Neighbor*), sendo semelhante ao IBL (VIJAYARANI; MUTHULAKSHMI, 2013, p. 3120), proposto por Aha, Kibler e Albert em 1991. As autoras esclarecem que, por meio do cálculo da distância euclidiana, IBL encontra a instância de treinamento mais próxima da instância de teste, atribuindo a mesma classe da instância de treinamento à instância de teste. Nesse algoritmo, caso várias instâncias sejam qualificadas como mais próximas, a primeira encontrada será utilizada (VIJAYARANI; MUTHULAKSHMI, 2013, p. 3120-3121).

Os algoritmos de aprendizagem bayesianos fundamentam-se no Teorema de Bayes para prever, estatisticamente, a probabilidade de um exemplo pertencer a determinada classe (CASTRO; FERRARI, 2016, p. 299; MITCHELL, 1997, p. 154). A regra de probabilidade foi introduzida pelo italiano Gerolamo Cardano e atualizada por Thomas Bayes (RUSSELL; NORVIG, 2013, p. 32) a qual, nos últimos tempos, tem

sido amplamente empregada na abordagem indutiva da inteligência artificial (BRAMER, 2016, p. 22; RUSSELL; NORVIG (2013, p. 32).

No algoritmo Naïve Bayes, frequentemente utilizado em experimentos para classificação de textos devido à sua simplicidade e eficácia (SCHNEIDER, 2005, p. 682), assume-se que o valor de um atributo, em uma determinada classe, é independente dos valores dos demais atributos na definição da classe (AWAD; KHANNA, 2015, p. 15; CASTRO; FERRARI, 2016, p. 300; RUSSELL; NORVIG, 2013, p. 932). Como, em muitos casos, essa suposição não é verdadeira, porém, objetiva simplificar cálculos (CASTRO; FERRARI, 2016, p. 300), o algoritmo é denominado ingênuo, no inglês naïve (GOLDSCHMIDT; PASSOS, 2005, p. 101). Em se tratando de classificação de textos, o algoritmo adota que não há dependência entre as palavras ou termos do documento (SCHNEIDER, 2005, p. 682).

Mitchell (1997, p. 182) descreve as etapas da classificação de documentos utilizando-se o classificador Naïve Bayes. Inicialmente o classificador examina todos os documentos de treinamento, com o propósito de extrair o conjunto de palavras ou vocabulário. Em seguida, contabiliza a frequência dessas palavras nas diferentes classes para, então, obter as estimativas das probabilidades. Na classificação de um documento não rotulado, o classificador utiliza essas probabilidades para calcular a classe mais provável do novo documento.

Há dois modelos distintos de classificadores indutivos que são comumente chamados de Naïve Bayes para classificar textos, o modelo binário e o multinomial (MCCALLUM; NIGAM, 1998, p. 41). Ambos calculam a probabilidade *a posteriori* de uma classe, tendo como base a distribuição das palavras no documento (AGGARWAL; ZHAI, 2012, p. 182).

No modelo binário o documento é representado por um vetor com indicação da presença ou ausência das palavras, e para o cálculo da probabilidade multiplicam-se as probabilidades de todos os valores atribuídos às palavras, incluindo a probabilidade da não ocorrência de palavras que não estão no documento (MCCALLUM; NIGAM, 1998, p. 41). Como geralmente são atribuídos os valores 0, para ausência e 1 para presença da palavra no documento (EYHERAMENDY; LEWIS; MADIGAN, 2003, 93), esse modelo descreve a distribuição baseada no modelo multivariado de Bernoulli (MCCALLUM; NIGAM, 1998, p. 41).

No modelo multinomial as palavras são representadas pelas frequências com que ocorrem no documento (EYHERAMENDY; LEWIS; MADIGAN, 2003, p. 95;

MCCALLUM; NIGAM, 1998, p. 41). Para calcular a probabilidade do documento, nesse caso, multiplica-se as probabilidades das palavras que apresentam alguma ocorrência (MCCALLUM; NIGAM, 1998, p. 41; AGGARWAL; ZHAI, 2012, p 182).

McCallum e Nigam (1998, p. 42), ao avaliarem esses dois modelos em seu estudo, concluíram que o multinomial supera o binário, principalmente quando o vocabulário inclui muitas palavras. Para Schneider (2005, p. 682), mesmo desconsiderando a forte dependência entre as palavras de um documento, como estrutura sintática e semântica, o Naïve Bayes apresenta um bom desempenho na classificação de textos. Witten *et al.* (2017, p. 103) definem a fórmula do algoritmo multinomial, que prevê a probabilidade do documento E pertencer à classe H já conhecida, por:

$$P(E|H) = N! \times \prod_{i=1}^{k} \frac{P_i^{ni}}{n_i!} \quad (1)$$

Onde:

$N = n_1 + n_2 + \ldots + n_k$ sendo n_1, $n_2, \ldots, n_k$ o número de vezes que a palavra ou termo i ocorre no documento;

P_1, $P_2, \ldots, P_k$ = probabilidade de se obter palavra ou termo i dados os documentos (exemplos) de treinamento na categoria H.

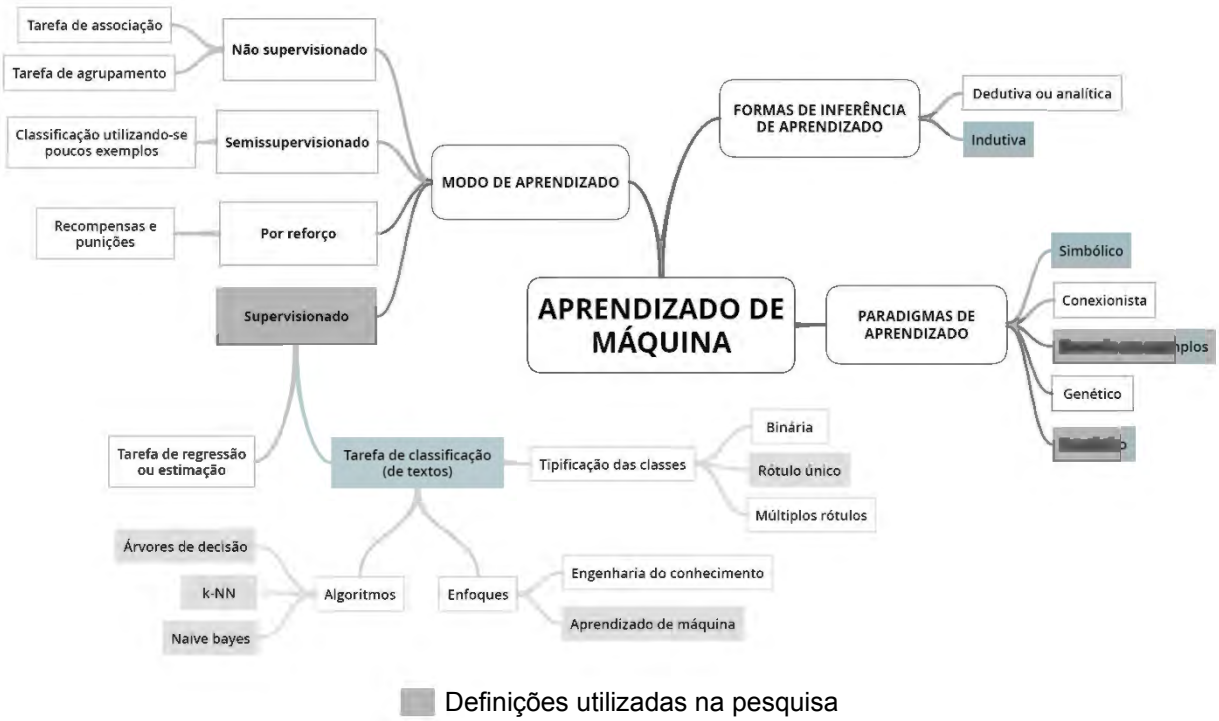
Para Witten *et al.* (2017, p. 103) a probabilidade do documento E pertencer à classe H é independente do contexto do termo e de sua posição no documento, e explicam que o uso de fatorial se dá justamente porque a ordem da ocorrência de cada termo é irrelevante no modelo *bag of words.*

Castro e Ferrari (2016, p. 299-300) afirmam que os algoritmos bayesianos apresentam alta acurácia e velocidade de processamento, inclusive em bases com grande número de exemplos. E por fim, os estudos apontados por Mitchell (1997, p. 154) sugerem que o classificador Naïve Bayes é competitivo e que, em alguns casos, supera outros algoritmos de aprendizagem, como as árvores de decisão e redes neurais. Castro e Ferrari (2016, p. 299-300) compartilham dessa mesma opinião.

Outros algoritmos utilizados e citados na literatura para classificação de textos baseiam-se em redes neurais artificiais e nas máquinas de vetores de suporte, *support vector machines* (SVM) no inglês. As redes neurais artificiais consistem em modelos matemáticos inspirados no funcionamento dos neurônios, com capacidade de adquirir e utilizar conhecimento experimental para aprendizado, sendo que os dados ficam distribuídos em uma estrutura de rede e são pesquisados de modo paralelo e não sequencial (GOLDSCHMIDT; PASSOS, 2005, p. 174). As máquinas de vetores de suporte determinam separadores, ou hiperplanos que melhor dividem as classes num espaço de busca (AGGARWAL; ZHAI, 2012, p. 194).

A Figura 5 sintetiza o conteúdo abordado neste tópico, o aprendizado de máquina.

FIGURA 5 – DIAGRAMA REFERENTE AO TÓPICO APRENDIZADO DE MÁQUINA



FONTE: A autora (2020) com auxílio do software GoConqr (EXAMTIME, 2021)

Para que os algoritmos de aprendizado possam ser aplicados a dados textuais faz-se necessário seu tratamento por meio de processamento de linguagem natural, tópico abordado a seguir.

2.4 PROCESSAMENTO DE LINGUAGEM NATURAL

O processamento de linguagem natural (PLN), no inglês *natural language processing* (NLP), compreende a área da inteligência artificial que integra sistemas informatizados capazes de analisar e sintetizar a linguagem falada ou escrita (JACKSON; MOULINIER, 2002, p. 2-3). A palavra “natural” é empregada com a finalidade de diferenciar a linguagem humana de outras linguagens, como aquelas voltadas à programação de computadores (JACKSON; MOULINIER, 2002, p. 2-3).

Liddy (2001, p. 2), em consenso a Jackson e Moulinier (2002, p. 2-3), afirma que o objetivo maior do PLN é o entendimento da linguagem natural, do mesmo modo que o humano o faz. Ademais, o termo inicialmente proposto para esse área foi “compreensão de linguagem natural”, no inglês *natural language understanding* (NLU), no sentido de realmente conseguir interpretar textos ou falas humanas, fazer inferências a partir de textos, responder a perguntas sobre o conteúdo de um texto e realizar traduções (LIDDY, 2001, p. 3). Para Gonzalez e Lima (2003, p. 349), o PLN

> trata computacionalmente os diversos aspectos da comunicação humana, como sons, palavras, sentenças e discursos, considerando formatos e referências, estruturas e significados, contextos e usos. Em sentido bem amplo, podemos dizer que o PLN visa fazer o computador se comunicar em linguagem humana, nem sempre necessariamente em todos os níveis de entendimento e/ou geração de sons, palavras, sentenças e discursos (GONZALEZ; LIMA, 2003, p. 349).

Esses níveis de entendimento são utilizados pelos humanos para transmitir e obter conhecimento e, portanto, quanto melhor for o sistema de PLN, mais níveis utilizará. Os níveis mais comuns são o fonético e fonológico, morfológico, sintático, semântico, pragmático (GONZALEZ; LIMA, 2003, p. 349; LIDDY, 2001, p. 7-9), o léxico e discurso (LIDDY, 2001, p. 7-9).

O nível fonético e fonológico trata do relacionamento entre as palavras e os sons que estas produzem (GONZALEZ; LIMA, 2003, p. 349), das variações de sons quando as palavras são pronunciadas em conjunto, da tonicidade e entonação (LIDDY, 2001, p. 7).

O morfológico atua nos morfemas, que são as menores unidades de significado das palavras, de modo a classificá-las em categorias morfológicas (GONZALEZ; LIMA, 2003, p. 349; LIDDY, 2001, p. 7), por exemplo gênero (masculino e feminino), número (singular e plural), grau (aumentativo, diminutivo, normal e

superlativo), tempo (futuro, passado e presente). As variações linguísticas, nesse nível, geralmente, abarcam mudanças das palavras considerando-se flexões, ou inflexões, e derivações. As flexões descrevem mudanças previsíveis que as palavras sofrem em função da sintaxe sem, contudo, haver alteração de classe gramatical, sendo as mais comuns o plural e os tempos verbais (ARAMPATZIS *et al.,* 2000, p. 5), citados anteriormente. Nas derivações, pode ou não ocorrer mudança gramatical, por exemplo, em duas palavras derivadas da raiz "computo", computar é verbo e computador é substantivo (ARAMPATZIS *et al.,* 2000, p. 6).

No nível léxico, cada palavra é associada a uma simples classe gramatical e aquelas palavras que possuem um único significado são substituídas pela representação semântica desse significado (LIDDY, 2001, p. 7). Para Gonzalez e Lima (2003, p. 350), o termo léxico significa "uma relação de palavras com suas categorias gramaticais e seus significados", sendo objetivo dos dicionários léxicos fornecer informações sobre as palavras como origem, pronúncia e sintaxe. Como a palavra é decomposta em unidades menores e, dada a existência de primitivas semânticas usadas em todas as palavras, o sistema de PLN consegue unificar o significado entre as palavras, da mesma forma que os humanos costumam fazer (LIDDY, 2001, p. 7). Esse processo é realizado por um etiquetador gramatical e semântico, no inglês *part-of-speech tagger* (GONZALEZ; LIMA, 2003, p. 358).

O nível sintático preocupa-se em analisar as palavras de cada frase com o propósito de identificar sua estrutura gramatical, a dependência entre as palavras (LIDDY, 2001, p. 8) e o relacionamento entre as frases, formando as sentenças (GONZALEZ; LIMA, 2003, p. 339).

No nível semântico busca-se identificar o relacionamento das palavras com seus possíveis significados, de modo a constituir o significado das sentenças (GONZALEZ; LIMA, 2003, p. 349; LIDDY, 2001, p. 8). Esse nível de processamento pode incluir desambiguação semântica de palavras com múltiplos sentidos (LIDDY, 2001, p. 8).

No discurso a preocupação é com unidades de texto, e não somente sentenças. Ao invés de trabalhar com sentenças individualmente, esse nível concentra-se nas propriedades do texto, de modo a transmitir um significado a partir do relacionamento existente entre as sentenças (LIDDY, 2001, p. 9). E o nível pragmático objetiva identificar o contexto subjacente de um texto, para além do significado identificado (LIDDY, 2001, p. 9).

O PLN tem sido utilizado em muitas aplicações, dentre essas a recuperação e a extração de informações, o resumo ou sumarização de textos e a classificação de textos.

A recuperação de informações visa ao desenvolvimento de algoritmos capazes de recuperar informações de repositórios de documentos, em especial, informações textuais. Nesse caso, o usuário faz uma consulta, inserindo o(s) termo(s) desejado(s) e o sistema retorna uma lista de documentos (MANNING; SCHÜTZE, 1999, p. 530), como nos buscadores da internet.

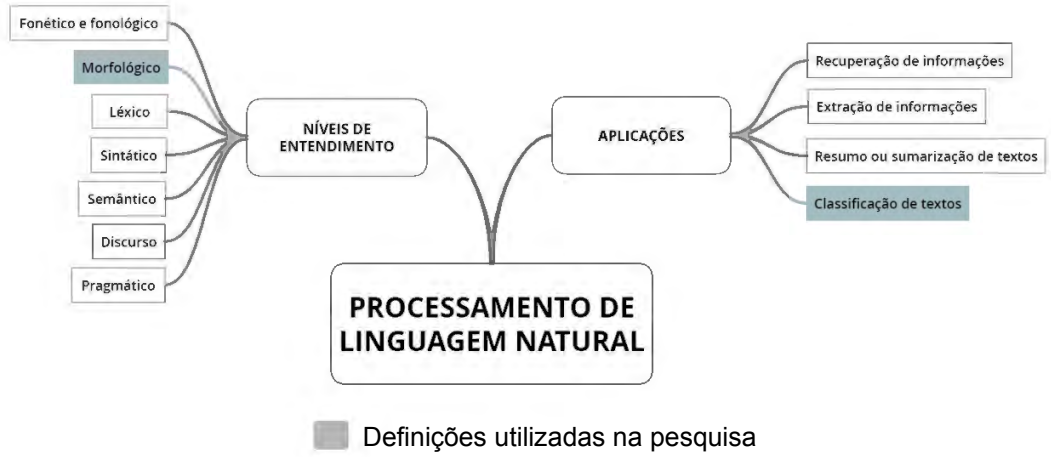
A extração de informações concentra-se no reconhecimento, na marcação e na extração de informações específicas em coleções de documentos, tais como pessoas, locais e empresas (LIDDY, 2001, p. 14) ou de determinados eventos, como desastres naturais, ataques terroristas e fusões de empresas, dentre muitos outros (MANNING; SCHÜTZE, 1999, p. 112).

A sumarização compreende o resumo de textos, que ocorre por meio da seleção automática de sentenças, parágrafos e, até mesmo, pela junção de parte ou de várias sentenças, gerando sentenças coerentes, de modo a resumir documentos (CHOWDHURY, 2003, p. 60). Para Liddy (2001, p. 14) a sumarização, no nível de discurso do PLN, permite a redução de textos, por meio de uma representação narrativa abreviada do documento original. O recurso de sumarização, geralmente, funciona adequadamente para pequenas coleções de documentos ou para documentos dentro de um domínio (CHOWDHURY, 2003, p. 60).

A classificação de textos foi discutida no tópico 2.3.3 desta pesquisa, bem como alguns dos algoritmos utilizados nessa tarefa, no âmbito de textos. Para Weiss, Indurkhya e Zhang (2015, p. 35), o processamento linguístico é parte relevante na classificação de textos ao atuar na identificação das características relevantes dos textos.

A Figura 6 sintetiza o conteúdo abordado neste tópico, o processamento de linguagem natural.

FIGURA 6 – DIAGRAMA REFERENTE AO TÓPICO PROCESSAMENTO DE LINGUAGEM NATURAL



FONTE: A autora (2020) com auxílio do software GoConqr (EXAMTIME, 2021)

Para que textos possam ser utilizados como insumo das tarefas de aprendizado de máquina, seu conteúdo precisa ser organizado em formato adequado ao processamento computacional. É deste assunto que trata o tópico seguinte.

2.5 REPRESENTAÇÃO DE DOCUMENTOS

O conjunto de documentos, ou *corpus*, é fundamental para que a classificação de textos possa ocorrer. Um *corpus* pode ser formado por qualquer agrupamento de documentos textuais, contudo, na maioria das vezes, agrega grandes coleções de documentos (FELDMAN; SANGER, 2007, p. 3).

Dada a enorme quantidade de palavras, os algoritmos envolvidos na classificação de textos necessitam de uma representação estruturada do *corpus,* para que possam ser executados. Nesse contexto, o espaço vetorial é um dos modelos mais populares de representação utilizado na tarefa de classificação. Feldman e Sanger (2007, p. 81), descrevem que nesse modelo, tradicionalmente,

> os documentos são representados por vetores de características. Uma característica é simplesmente uma entidade desprovida de estrutura interna – uma dimensão no espaço de características. Um documento é representado como um vetor nesse espaço – uma sequência de características e seus pesos. O modelo mais comum de saco de palavras simplesmente utiliza todas as palavras de um documento como características e, portanto, a dimensão do espaço de características é igual ao número de diferentes palavras em todos os documentos (FELDMAN; SANGER, 2007, p. 81, tradução nossa).

Rossi (2015, p. 15) complementa essa definição ao esclarecer que o modelo saco de palavras, ou *bag-of-words* no inglês (BOW), representa a coleção de documentos por meio de vetores, que são associados a cada texto, e dimensões, que correspondem aos termos ou características dos documentos. Nesse modelo, também denominado matriz de termos e documentos, no inglês *term document matrix* (TDM / DTM) (MANNING; RAGHAVAN; SCHÜTZE, 2009, p. 123), ou representação por atributo-valor (MONARD; BARANAUSKAS, 2003, p. 45), um termo pode ser representado por palavras simples, unigramas, ou por palavras compostas, denominadas bigramas, trigramas, n-gramas, que ocorrem no documento (MARTINS; MONARD; MATSUBARA, 2003, p. 229; TAN; WANG; LEE, 2002, p. 532-536). O modelo *bag-of-words* é exposto na Tabela 1.

TABELA 1 – REPRESENTAÇÃO DE DOCUMENTOS NO MODELO *BAG-OF-WORDS*

Documento / Termo	A_1	A_2	...	A_m	C
E_1	A_{11}	A_{12}	...	A_{1m}	C_1
E_2	A_{21}	A_{22}	...	A_{2m}	C_2
...	...	...	...	...	...
E_n	A_{n1}	A_{n2}	...	A_{nm}	C_n

FONTE: A autora (2020) com base em Martins, Monard e Matsubara (2003, p. 229)

Observa-se que existem n documentos, ou exemplos no caso do conjunto de treinamento, e m atributos ou termos. Cada documento (E) é composto por atributos (A), ficando implícito que (E) compreende um vetor, no qual a última dimensão abrange a classe (C) à qual pertence o documento.

Silva e Souza (2014, p. 2) esclarecem que n-gramas expressam a dependência entre as palavras de um *corpus*, a qual é mensurada a partir das frequências das coocorrências dessas palavras, isto é, quando aparecem juntas. Em vista disso, termos formados por duas palavras adjacentes $(n = 2)$ são denominados bigramas.

O uso de bigramas e trigramas, para Russell e Norvig (2013, p. 998), eleva o número de características ao quadrado ou ao cubo e para Tan, Wang e Lee (2002, p. 533) os termos formados por n-gramas maiores que três tendem a diminuir o desempenho dos classificadores. De modo oposto, para Han, Karypis e Kumar (1999,

p. 383) palavras isoladas tendem a não qualificar a categoria de documentos, contudo, palavras dependentes, como bigramas ou trigramas, podem determinar com exclusividade a categoria correta.

O *bag-of-words* desconsidera a ordem, o contexto e a relação entre as palavras ou termos dos textos (WITTEN, 2004, p. 7; ROSSI, 2015, p. 19; RUSSELL; NORVIG, 2013, p. 998). Essas questões envolvem, essencialmente, os níveis de entendimento léxico, semântico e de discurso que, em geral, são tratados por ontologia.

Ontologias são descritas como o conjunto de todas as classes de interesse e todas as relações entre essas classes para um determinado domínio (FELDMAN; SANGER, 2007, p. 42), constituindo, dessa maneira, uma base ampla de conhecimento do domínio (RUSSELL; NORVIG, 2013, p. 1013).

A opção pela utilização do *bag-of-words* baseia-se na afirmação de Weiss, Indurkhya e Zhang (2015, p. 8). Para os autores, muitas aplicações que utilizam esse modelo apresentam resultados bem sucedidos, embora o modelo seja simples. Igualmente na menção de Aggarwal e Zhai (2012, p. 3), quando apontam que, mesmo havendo perda de informações sobre o posicionamento dos termos no *bag-of-words*, a maioria das aplicações de classificação de textos o utilizam, devido à simplicidade.

2.5.1 Ponderação de termos

Constituído o modelo com os termos e documentos do *corpus* sob análise, é necessário estipular pesos para as características presentes, com vistas à aplicação dos algoritmos. Normalmente os termos apresentam diferentes graus de importância no conjunto de documentos. As abordagens mais usuais para atribuição de pesos às características são a booleana, ou binária, a frequência do termo, *term frequency* (TF) no inglês, e a frequência do termo – frequência inversa no documento, *term frequency - inverse document frequency* (TF-IDF) no inglês.

A abordagem binária é a mais simples e considera a presença ou ausência dos termos em cada documento, atribuindo 0 para ausência e 1 para presença (FELDMAN; SANGER, 2007, p. 68; SEBASTIANI, 2002, p. 11). Como nessa abordagem os termos presentes possuem a mesma importância, torna-se inadequada para muitas aplicações. Nesse sentido, as medidas de ponderação TF e TF-IDF consideram a frequência com que um termo aparece num documento e a frequência

com que esse termo é encontrado em outros documentos, e podem apresentar melhores resultados (MARTINS; MONARD; MATZUBARA, 2003, p. 229).

A frequência do termo considera o número de ocorrências do termo no documento (JACKSON; MOULINIER, 2002, p. 34; MARTINS; MONARD; MATZUBARA, 2003, p. 229; MANNING; RAGHAVAN; SCHÜTZE, 2009, p. 117), por exemplo, se o termo aparece 5 vezes no documento, então esse valor é armazenado como atributo no vetor e dimensão correspondentes. Termos que aparecem com maior frequência, isto é, em quase todos os documentos, podem não ser úteis para discriminar os documentos. Nesse sentido, a medida TF-IDF favorece termos que ocorrem em poucos documentos (MARTINS; MONARD; MATZUBARA, 2003, p. 229).

Na abordagem TF-IDF considera-se que IDF é inversamente proporcional ao número de documentos com um determinado termo numa coleção de documentos (MANNING; RAGHAVAN; SCHÜTZE, 2009, p. 118; MARTINS; MONARD; MATZUBARA, 2003, p. 229), portanto, um termo que aparece em todos os documentos do *corpus* terá IDF igual a zero (JACKSON; MOULINIER, 2002, p. 35). Diante disso, a ponderação TF-IDF constitui-se da combinação das definições da frequência do termo e da frequência inversa no documento, de modo a produzir um peso composto para cada termo em cada documento (MANNING; RAGHAVAN; SCHÜTZE, 2009, p. 118).

2.5.2 Redução de dimensionalidade

No aprendizado de máquina são necessárias muitas características para treinar os classificadores, o que leva ao aumento da complexidade e de requisitos computacionais, bem como do tempo para processamento (AWAD; KHANNA, 2015, p. 29). No modelo *bag-of-words,* a representação das características tende a gerar atributos esparsos e matrizes altamente dimensionais, ou seja, muitos atributos que ocorrem em pouquíssimos documentos. Há uma diversidade de técnicas descritas na literatura que objetivam atenuar esse problema.

A conflação consiste em igualar as variantes morfológicas e diminuir a diversidade de palavras encontradas em um *corpus*. As técnicas mais conhecidas de conflação são a radicalização, ou *stemming*, e a redução à forma canônica, ou *lemmatization*. Nesse entendimento, para Manning, Raghavan e Schütze (2009, p.

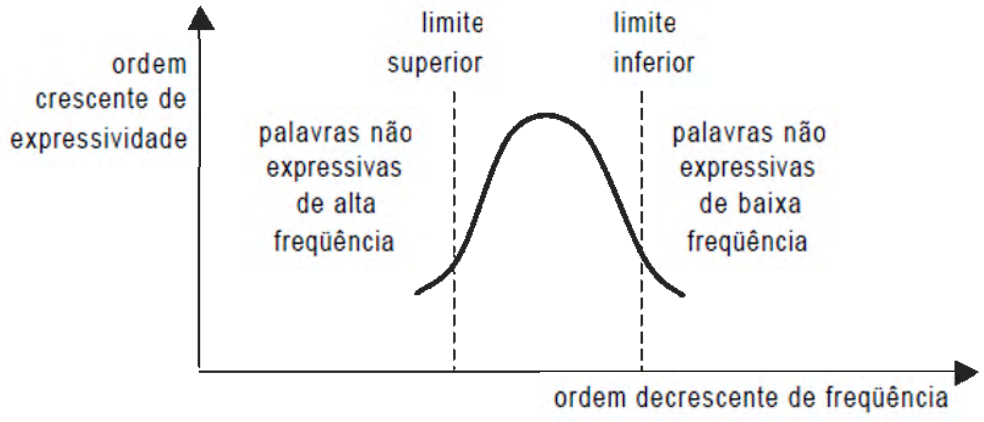
32), as técnicas de *stemming* e *lemmatization* objetivam reduzir as flexões e derivações de uma palavra.

O *stemming* consiste na redução das palavras com o mesmo radical a uma forma chamada *stem* (ARAMPATZIS *et al.,*2000, p. 5*;* MANNING; SCHÜTZE, 1999, p. 132), assumindo que todas as palavras com mesmo *stem* apresentam significado semelhante. Para tanto, há remoção dos afixos das palavras, geralmente prefixos e sufixos (ARAMPATZIS *et al.*, 2000, p. 5).

Com *lemmatization* os verbos são reduzidos ao infinitivo e os adjetivos e substantivos à forma no masculino singular (SANTOS, 2019, p. 23). Na classificação de textos, Weiss, Indurkhya e Zhang (2015, p. 18) salientam que *stemming* e *lemmatization* fornecem resultados modestos ao reduzir o número de palavras e aumentar a frequência de ocorrência de palavras individuais.

Zipf, em 1949, constatou a existência da frequência de ocorrência de palavras num documento, sendo possível ordená-las conforme essa frequência. Em 1959, Luhn sugeriu que as palavras com maiores frequências de ocorrência são pouco expressivas, pois condizem a artigos e preposições, da mesma forma que as palavras com baixas frequências, uma vez que raramente aparecem no conjunto de textos e, assim, formulou o Gráfico de Luhn, conforme a Figura 7 (GONZALEZ; LIMA, 2003, p. 359).

FIGURA 7 – GRÁFICO DE LUHN RELACIONANDO A EXPRESSIVIDADE E FREQUÊNCIA DE OCORRÊNCIA DAS PALAVRAS



FONTE: Gonzalez e Lima (2003, p. 359)

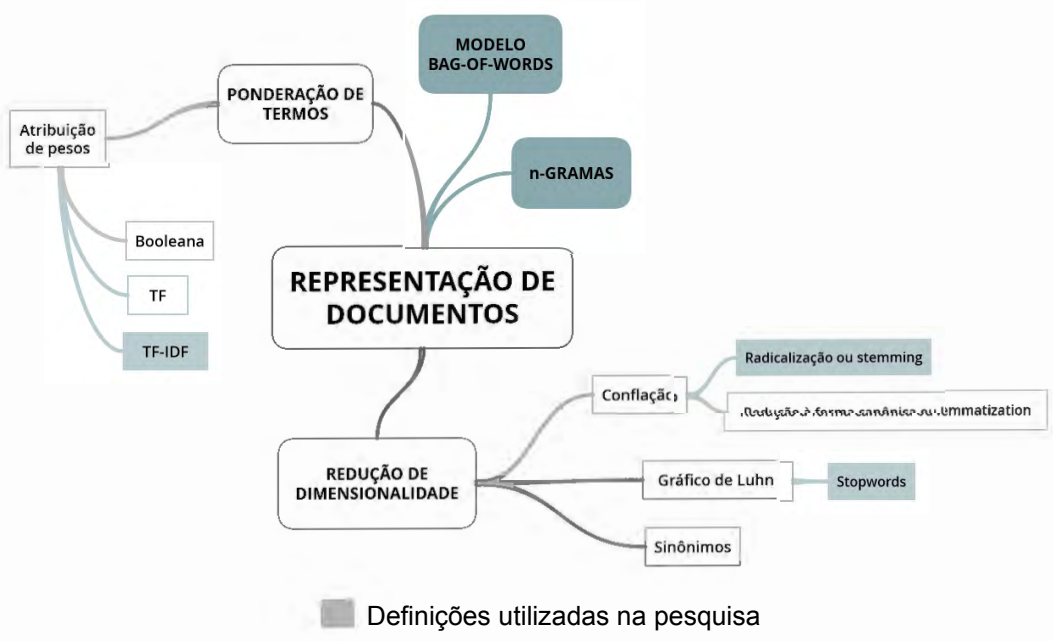
O gráfico demonstra que as palavras com frequência de ocorrência intermediária, entre os limites superior e inferior, são as expressivas e, portanto, constituem o vocabulário a ser considerado na classificação.

Nesse contexto, pode ser aplicada a técnica de remoção de palavras e de caracteres com grande incidência e que apresentam pouca relevância nos documentos, as chamadas *stopwords* (FELDMAN; SANGER, 2007, p. 6; MANNING; SCHÜTZE, 1999, p. 533; WEISS; INDURKHYA; ZHANG, 2015, p. 22). *Stopwords* compõem uma lista de palavras, geralmente, formada por preposições, artigos, pronomes e caracteres especiais que não fornecem qualquer potencial preditivo. Para Manning e Schütze (1999, p. 533) *stopwords* são palavras irrelevantes que podem ser ignoradas, quando a análise do texto envolve características que compreendem palavras-chave.

A utilização de sinônimos (ARAMPATZIS *et al.,* 2000, p. 5) também é uma técnica utilizada para diminuir a pluralidade de palavras e melhorar o desempenho das tarefas do aprendizado de máquina. No entanto, os autores ressaltam que existe o problema de ambiguidade das palavras, que também precisa ser tratado.

A Figura 8 sintetiza o conteúdo abordado neste tópico, a representação de documentos.

FIGURA 8 – DIAGRAMA REFERENTE AO TÓPICO REPRESENTAÇÃO DE DOCUMENTOS



FONTE: A autora (2020) com auxílio do software GoConqr (EXAMTIME, 2021)

As técnicas de processamento de linguagem natural e de representação dos documentos no modelo *bag-of-words* podem ser viabilizadas por meio de ferramentas da mineração de textos, assunto discutido no próximo tópico.

2.6 MINERAÇÃO DE TEXTOS

A mineração de textos, *text mining* no inglês, consiste num processo intensivo de conhecimento, no qual são utilizadas ferramentas para análise de coleções de documentos (FELDMAN; SANGER, 2007, p. 1). Semelhante à mineração de dados, que procura descobrir padrões em dados estruturados, a mineração de textos procura tratar e extrair informações relevantes de textos, segundo determinado propósito (WEISS; INDURKHYA; ZHANG, 2015, p. 1; WITTEN *et al.*, 2017, p. 515). Para Aranha e Passos (2006, p. 1) a mineração de textos, também chamada de mineração de dados textuais ou descoberta de conhecimento de bases de dados textuais, é

> um campo novo e multidisciplinar que inclui conhecimentos de áreas como Informática, Estatística, Linguística e Ciência Cognitiva. Mineração de textos consiste em extrair regularidades, padrões ou tendências de grandes volumes de textos em linguagem natural, normalmente, para objetivos específicos. Inspirado pelo data mining ou mineração de dados, que procura descobrir padrões emergentes de banco de dados estruturados, a mineração de textos pretende extrair conhecimentos úteis de dados não estruturados ou semiestruturados (ARANHA; PASSOS, 2006, p. 1).

Feldman e Sanger (2007, p. x) esclarecem que a mineração de textos procura resolver a crise de sobrecarga de informações, por meio da combinação das áreas de mineração de dados, aprendizado de máquina, processamento de linguagem natural, recuperação de informações e gestão do conhecimento. A mineração de textos utiliza as mesmas técnicas da mineração de dados, como classificação, agrupamento, associação e técnicas voltadas exclusivamente ao processamento de dados textuais, entre estas, sumarização e análise de diferenças e similaridades entre textos (LOH, 2014, p. 149).

Quando utiliza algum mecanismo de busca, o usuário já sabe o que deseja, ao contrário da mineração, que visa à descoberta de informações desconhecidas (ARANHA; PASSOS, 2006, p. 2), por meio da identificação de padrões e informações discrepantes em textos (AGGARWAL; ZHAI, 2012, p. 2). Textos desestruturados são transformados num formato de planilha, na qual as linhas são exemplos de

experiências anteriores e as colunas constituem seus atributos, dessa maneira, um documento pode ser considerado um exemplo completo para treinamento (WEISS; INDURKHYA; ZHANG, 2015, p. 3).

Aranha e Passos (2006, p.4) dividem o processo de mineração de textos em cinco etapas: coleta, pré-processamento, indexação, mineração e análise. A coleta consiste na composição do conjunto de documentos a ser trabalhado. Compreende, geralmente, uma atividade penosa, em razão dos documentos estarem dispersos, em formatos distintos, ou devido à dificuldade de encontrá-los e selecioná-los nos repositórios A internet é um exemplo típico desses problemas, sendo necessário o uso de *crawlers*, softwares robôs responsáveis pela coleta automatizada dos textos (ARANHA, 2007, p. 43-44).

O pré-processamento consiste em preparar, por meio de transformações, o conjunto de textos para que possa ser representado num modelo como o *bag-of-words*. As transformações, por vezes, realizam a “limpeza” dos textos, uma vez que compreendem a aplicação das técnicas de *stemming*, *lemmanization*, remoção das *stopwords* e de caracteres especiais, números, *links* etc, dependendo dos resultados a serem alcançados.

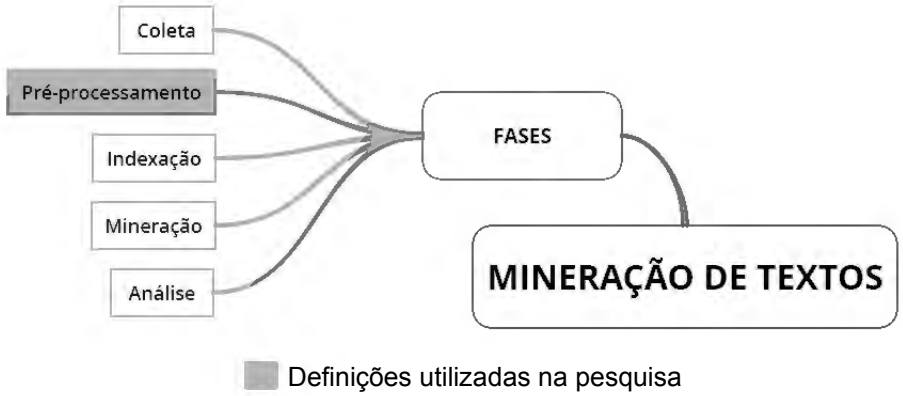
A etapa de pré-processamento também engloba a representação dos textos num formato adequado para a mineração. Esse procedimento consiste em dividir cada texto em partes menores, geralmente palavras ou termos, de acordo com a especificidade linguística (ARANHA; PASSOS, 2006, p. 3).

A indexação corresponde à aplicação de técnicas que permitem a busca rápida, por palavras-chave ou outros símbolos que consideram aspectos semânticos, permitindo a leitura e manipulação eficiente de grandes volumes de textos (ARANHA, 2007, p. 50-51).

A etapa de mineração envolve a decisão de quais algoritmos deverão ser aplicados, os quais são provenientes de áreas como aprendizado de máquina, estatística e bancos de dados (ARANHA, 2007, p. 60). E, por fim, a etapa de análise de dados, que envolve avaliação e interpretação dos resultados providos pelos algoritmos, por meio da utilização de métricas, como taxas de erro e desempenho (ARANHA, 2007, p. 61).

A Figura 9 sintetiza o conteúdo abordado neste tópico, a mineração de textos.

FIGURA 9 – DIAGRAMA REFERENTE AO TÓPICO MINERAÇÃO DE TEXTOS



FONTE: A autora (2020) com auxílio do software GoConqr (EXAMTIME, 2021)

Tendo em vista a finalização desta revisão de literatura, torna-se adequado trazer alguns dos trabalhos que auxiliaram na definição dos tópicos discutidos.

2.7 TRABALHOS RELACIONADOS

Neste tópico são apresentados alguns dos trabalhos relacionados à pesquisa, seja pela descrição de benefícios ou problemas quanto à utilização de aprendizado de máquina no âmbito governamental, seja pela abrangência dos tópicos discutidos na revisão de literatura e voltados aos serviços públicos, emergenciais e no atendimento ao cidadão ou consumidor de serviços. O Quadro 2 traz uma compilação desses trabalhos e dos conceitos abordados.

O trabalho de Andrade (2019) apresenta uma proposta para inovação na gestão das informações da Companhia do Desenvolvimento do Planalto Central – CODEPLAN. O objetivo é a integração das bases de dados acerca da prestação de serviços públicos do Governo do Distrito Federal. Para o autor, a associação entre os dados registrados permite tomadas de decisões no nível estratégico, de modo a direcionar ações governamentais para as reais necessidades dos cidadãos. As fases do trabalho compreenderam: planejamento, contemplando o mapeamento dos dados governamentais existentes; identificação de informações executivas; mapeamento das bases fonte; definição dos metadados; definição da base de dados e; elaboração de um modelo multidimensional, com a finalidade de análise de dados. Dentre os resultados esperados estão a disponibilização de informações corretas e precisas; a realização de novas análises; a formulação de propostas para melhoria de processos;

o monitoramento das demandas da população e da organização; diminuição de retrabalho e; viabilização de indicadores de políticas públicas.

QUADRO 2 – ALGUNS DOS TRABALHOS RELACIONADOS À PESQUISA E AOS CONCEITOS ABORDADOS

Conceito	Autor (ano) - tipo de documento	Título
Gestão da informação na administração pública	Andrade (2019) - tese	Proposta de inovação na gestão da informação na Companhia de Planejamento do Distrito Federal – CODEPLAN.
Gestão da informação na administração pública	Condurú e Pereira (2017) - artigo	Gestão da informação em saneamento básico no Estado do Pará sob o enfoque do ciclo informacional.
Gestão da informação no terceiro setor	Rodrigues e Parrão (2017) - artigo	A necessidade da implantação da gestão da informação no ADRA/CADECA.
Aprendizado de máquina, PLN e mineração de textos	Androutsopoulou *et al.* (2019) - artigo	*Transforming the communication between citizens and government through AI-guided chatbots.*
Inteligência artificial e aprendizado de máquina	Kuziemski e Misuraca (2020) - artigo	*AI governance in the public sector: Three tales from the frontiers of automated decision-making in democratic settings.*
Inteligência artificial e aprendizado de máquina	Mehr, Ash e Fellow (2017) - *working paper*	*Artificial Intelligence for Citizen Services and Government.*
Inteligência artificial, aprendizado de máquina; classificação de textos, algoritmos para classificação de textos; representação de documentos	Pollettini (2016) - tese	Avaliação de mecanismos de suporte à tomada de decisão e sua aplicabilidade no auxílio à priorização de casos em regulações de urgências e emergências.
PLN, classificação de textos, algoritmos para classificação de textos; representação de documentos	Berti (2017) - tese	Modelo preditivo de situações como apoio à consciência situacional e ao processo decisório em sistemas de resposta à emergência.
Aprendizado de máquina, PLN, classificação de textos e algoritmos para classificação de textos	Baghdadi *et al.* (2019) - artigo	*Automatic classification of free-text medical causes from death certificates for reactive mortality surveillance in France.*
Mineração e classificação de textos	Kano, Fujita e Tsuda (2019) - artigo	*A Method of Extracting and Classifying Local Community Problems from Citizen-Report Data using Text Mining.*
Mineração de textos	Monteiro (2017) - tese	Mensagens textuais no canal de atendimento do portal ibgeando: obtendo insumos para a tomada de decisão utilizando mineração de textos.

FONTE: A autora (2021)

Condurú e Pereira (2017) ressaltam que a informação representa um recurso estratégico no planejamento e desenvolvimento de municípios, sendo fundamental para identificação das demandas da sociedade. A pesquisa concentra-se no processo

de gestão da informação da área de saneamento básico do Estado do Pará, na qual observa-se falta de integração entre dados, conduzindo a informações divergentes e ocasionando a elaboração de políticas públicas nem sempre compatíveis com as reais necessidades dos cidadãos. O estudo fundamenta-se no Ciclo de Transferência de Informação de Lancaster, que compreende quatro fases. Na primeira, a produção e registro de dados e informações, os autores destacam que o registro da informação ocorre de modo fragmentado e desintegrado, acarretando prejuízos na tomada de decisão pelo Estado. Na segunda fase, referente à sistematização das informações, apontam que há falta de detalhamento do relacionamento de informações e de alinhamento entre órgãos envolvidos no saneamento da região. Na terceira fase, que contempla a disseminação dos dados e informações organizados, os autores destacam a necessidade das informações sistematizadas da etapa anterior, além da periodicidade de atualização. Consequentemente, na última fase, relativa ao uso de informações, os autores constatam o uso de bases do governo federal, haja vista a falta de bases locais para diagnóstico, planejamento e monitoramento de ações; elaboração de projetos de engenharia; captação de recursos e; definição de investimentos.

A pesquisa organizada por Rodrigues e Parrão (2017) objetivou iniciar o processo de gestão da informação na instituição ADRA/CADECA, em Presidente Prudente – São Paulo. Ofertando serviços de convivência e fortalecimento de vínculos, a instituição atende crianças e adolescentes, de seis a 14 anos, em situação de vulnerabilidade e risco social. Com a implantação de um banco de dados pretende-se registrar dados de identificação, escolarização, situação habitacional, saúde e familiares da população atendida. As autoras esclarecem que, devido à criação da Secretaria de Avaliação e Gestão da Informação no Ministério de Desenvolvimento Social, houve avanço na área da assistência social no que diz respeito ao monitoramento do repasse de recursos financeiros para estados e municípios e ao gerenciamento e à análise de processos que envolvem cidadãos em situação de vulnerabilidade e risco social. Com vistas a acompanhar o avanço observado na área governamental, para as autoras, é essencial a implantação de uma ferramenta que possibilite otimizar tempo e auxiliar a tomada de decisão dos profissionais da instituição, no atendimento das crianças e dos adolescentes.

A pesquisa de Androutsopoulou *et al.* (2019) apresenta uma abordagem que utiliza processamento de linguagem natural, aprendizado de máquina e tecnologias

de mineração de dados para aprimoramento do relacionamento entre governo e cidadãos na Grécia. *Chatbots* utilizam dados estruturados e semanticamente organizados para interagir com os cidadãos. As técnicas propiciam, desde a coleta de dados, em diferentes formatos e fontes, até a distribuição de informações. Os algoritmos utilizados baseiam-se em redes neurais, árvores de decisão, SVM e Naïve Bayes. Os autores apontam que o estudo contribui com as pesquisas de IA realizadas para o setor público, que objetivam aprimorar a comunicação entre governos e cidadãos, que há grande potencial de utilização dos *chatbots*, contudo, que o sucesso depende da construção de bases de conhecimento sólidas.

O estudo conduzido por Kuziemski e Misuraca (2020) procurou investigar como o uso do aprendizado de máquina no setor público pode estar intensificando as desigualdades de poder entre governos e cidadãos. Foram analisados os serviços de controle de imigração no Canadá, de empregos na Polônia e de experiência digital na Finlândia. A pesquisa baseou-se na análise das implicações legais do uso da IA e na avaliação dos métodos de governança de dados de IA que fortalecem a legitimidade do governo. Os autores concluem que mesmo a aplicação simples dos algoritmos pode se tornar um instrumento controle sobre os cidadãos, ao fazer julgamentos de valor e acumular dados confidenciais.

O estudo apresentado por Mehr, Ash e Fellow (2017), de modo oposto, relaciona os benefícios que podem ser alcançados, tanto para governos quanto para cidadãos, com a utilização da IA no âmbito governamental. Os autores destacam o aumento da eficiência, a consonância com serviços e ferramentas digitais emergentes, já utilizados e disponibilizados pelo setor privado, auxílio aos servidores, agilização do atendimento e redução de encargos administrativos, entre outros. Os autores concluem que a inteligência artificial, incluindo a aprendizagem computacional, possui potencial para melhoria da interação governo-cidadão.

A pesquisa de Pollettini (2016) objetivou proporcionar suporte à decisão na definição de prioridades de casos médicos admitidos na Unidade de Emergência do Hospital das Clínicas da Faculdade de Medicina, em Ribeirão Preto, São Paulo (HCFMRP–USP). Inicialmente, dados das descrições clínicas e do diagnóstico dos pacientes foram processados para reconhecimento de conceitos, utilizando-se vocabulário médico controlado (metatesauros). Na sequência foi realizada definição de prioridade do caso, por meio de algoritmo análogo ao k-NN, com cálculos de similaridade realizados a partir de uma matriz de frequência de termos e documentos,

com métrica TF-IDF. Para dirimir a “maldição da dimensionalidade” da matriz utilizou-se *stemming* e correção ortográfica. Na classificação usou-se os algoritmos IBk, J48, RandomForest, RBFNetwork, MultilayerPerceptron, BayesNet, NaiveBayes e Vote, no software Weka.

A pesquisa de Berti (2017) buscou o desenvolvimento de mecanismos de apoio à tomada de decisão dos operadores de sistemas de resposta à emergência, especificamente, das chamadas do número 190, da Polícia Militar de São Paulo, integrado ao Corpo de Bombeiros e Samu. O modelo preditivo proposto prepara os dados textuais recebidos, identificando e classificando-os conforme situações preexistentes, disponibilizando ao operador um ranking das situações com maiores escores de classificação. Para orientar a classificação foram utilizadas descrições de 54 fluxogramas, comparando-se palavras desses fluxogramas às situações registradas. Utilizou-se o modelo *bag-of-words* e o classificador Naïve Bayes. Foram realizados testes com apenas 10 palavras, atingindo resultados bastante satisfatórios.

O estudo de Baghdadi *et al.* (2019) versa sobre a implantação e avaliação de desempenho de um modelo, baseado em aprendizado de máquina e processamento de linguagem natural, que objetiva monitorar e prever o impacto de doenças na população da França. O modelo procura classificar as causas de morte diagnosticadas pelos médicos e registradas nas certidões de óbito, muitas vezes com palavras incorretas. Estatísticas sobre esses dados levam entre seis e 24 meses para serem publicadas. O modelo apresentou alto desempenho, vindo a contribuir para a vigilância de mortalidade no país.

A pesquisa de Kano, Fujita e Tsuda (2019) acerca das demandas registradas no Sistema Relatório do Cidadão, em Chiba, no Japão, propôs um método para extrair, classificar e esclarecer a tendência dos problemas da cidade, de modo que a metodologia pudesse ser reproduzida para outros governos locais japoneses. Foi realizada análise morfológica dos textos por meio de processamento de linguagem natural, com agrupamento de termos adjacentes. Uma amostra foi classificada manualmente e, posteriormente, uma massa de registros submetida à classificação automática. O estudo demonstrou que danos em vias, que pareciam ser o principal problema relatado, apresentavam valor semelhante aos demais, cerca de 16%.

O trabalho de Monteiro (2017) procurou identificar, por meio da ferramenta SOBEK *Mining*, palavras condizentes às maiores dúvidas dos servidores do Instituto Brasileiro de Geografia e Estatística (IBGE), registradas no Portal IBGEANDO, para

as 20 categorias existentes. O objetivo foi obter insumos para auxiliar na tomada de decisão e melhorar a comunicação da área de Recursos Humanos.

Além de auxiliar na estruturação da revisão de literatura, esses trabalhos proporcionaram a identificação dos principais autores, para cada um dos assuntos que compõem a revisão de literatura. A partir dos trabalhos, também foi possível verificar como governos têm utilizado a tecnologia da informação para classificar as demandas dos cidadãos.

3 ENCAMINHAMENTOS METODOLÓGICOS

Este capítulo descreve a caracterização da pesquisa, bem como os procedimentos empregados para alcance dos objetivos propostos.

3.1 CARACTERIZAÇÃO DA PESQUISA

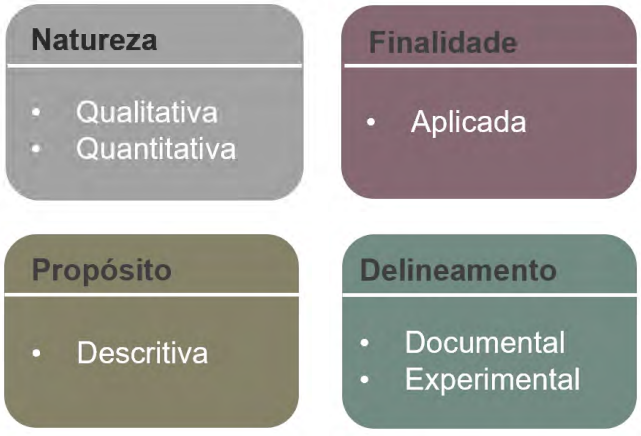
Quanto à natureza, a pesquisa se caracteriza como quantitativa, uma vez que utiliza quantificação, tanto para coleta quanto para tratamento dos dados (RICHARDSON, 2012, p. 70). É também qualitativa ao possibilitar o entendimento de um fenômeno social, permitindo descrever e analisar um problema (RICHARDSON, 2012, p. 79-80).

Quanto à finalidade, a pesquisa é aplicada, tendo em vista que utiliza técnicas de aprendizado supervisionado, processamento de linguagem natural e mineração de textos no conjunto amostral de protocolos da Central 156.

Quanto ao propósito, a pesquisa é classificada como descritiva, pois procura descobrir e sistematizar as relações entre variáveis de determinada população (GIL, 2008, p. 28), as quais influenciam, ou definem, a ocorrência de determinado fenômeno (RICHARDSON, 2012, p. 71). No presente estudo, as variáveis independentes compreendem aos termos e às palavras-chave dos documentos, também chamados de características dos textos, e as variáveis de controle correspondem às classes dos protocolos.

Quanto ao delineamento, a pesquisa caracteriza-se como documental e experimental. Documental devido à utilização de fontes documentais que não receberam tratamento analítico (GIL, 2008, p. 51), isto é, que não foram divulgadas em publicações científicas. Experimental uma vez que há manipulação variáveis independentes, ou causas, e, posteriormente, observação do efeito dessa manipulação em uma ou mais variáveis dependentes, ou de controle (RICHARDSON, 2012, p. 74; BARROS; LEHFELD, 2007, p. 91). É notório observar o efeito de uma ou mais variáveis em um único experimento, bem como as inter-relações e graus de intensidade e de influência de umas sobre as outras BARROS; LEHFELD, 2007, p. 91). Na Figura 10 é apresentada a síntese da caracterização da pesquisa.

FIGURA 10 – SÍNTESE DA CARACTERIZAÇÃO DA PESQUISA



FONTE: A autora (2021)

Posterior à caracterização, segue-se para descrição dos procedimentos empregados na pesquisa.

3.2 MATERIAIS E MÉTODOS

Neste tópico são descritos os procedimentos realizados na elaboração da pesquisa, compreendendo o levantamento bibliográfico, a pesquisa documental e a pesquisa experimental.

3.2.1 Levantamento bibliográfico

Em julho de 2019 foi realizado o levantamento bibliográfico visando justificar a elaboração da pesquisa bem como identificar os principais autores, aprofundar conceitos e compreender os diferentes enfoques acerca dos temas que fundamentam a pesquisa. Foram utilizados o Portal da Capes[3] e as bases Ebsco[4], Emerald Insight[5] e Science Direct[6], escolhidos por abarcar publicações relacionadas à tecnologia e à ciência da informação.

---

[3] Disponível em: https://www-periodicos-capes-gov-br.ez22.periodicos.capes.gov.br/index.php?. Acesso em 11 nov. 2019.

[4] Disponível em https://www-periodicos-capes-gov-br.ez22.periodicos.capes.gov.br/index.php?option=com_pmetabusca&mn=70&smn=78&base=find-db-1&type=b&Itemid=126. Acesso em 11 nov. 2019.

[5] Disponível em: https://www.emerald.com/insight/. Acesso em 11 nov. 2019.

[6] Disponível em: https://www.sciencedirect.com/. Acesso em 11 nov. 2019.

O levantamento foi realizado em duas etapas, considerando-se o período compreendido entre os anos 1965 e 2019 e para os idiomas português e inglês, conforme apresentado na Tabela 2.

TABELA 2 – EXPRESSÕES DE BUSCA, PORTAL E BASES DE DADOS UTILIZADOS NA PESQUISA DE PUBLICAÇÕES CIENTÍFICAS

Expressão de busca	Portal e bases de dados				
	Portal da Capes	Ebsco	Emerald	Science Direct	Total
(“aprendiza* de máquina” OR “aprendizado automático” OR “sistema* inteligente*” OR "inteligência computacional") AND (“classificação de texto*” OR “categorização de texto*” OR “mineração de texto*” OR “processamento de texto*” OR “processamento de linguagem natural”)	19	2	0	2	23
(“machine learning” OR “computer learning” OR “explanation-based learning” “learning classifier system*” OR “data mining” OR “automatic classification” OR “supervised learning”) AND (“text classification” OR “text mining” OR “text processing” OR “predictive text” OR “natural language processing”)	36.878	8.476	516	12.587	58.344
Total	36.897	8.478	516	12.589	58.480
Busca focada					
AND (“serviços públicos” OR “serviços municipais” OR “serviços emergenciais" OR "serviços ao cidadão" OR "serviços governamentais OR “atendimento ao cidadão”)	0	0	0	0	0
AND ("public service*" OR "municipal service*" OR "civil service*" OR "citizen service*" OR ("communit* service*" OR "govern* service*" OR "public utilit*" OR "emergency service*")	292	25	53	134	504
Total busca focada	292	25	53	134	504
Portal da Capes – Coordenação de Aperfeiçoamento de Pessoal de Nível Superior incluindo as bases de dados *Advanced Technologies & Aerospace Database; Advanced Technologies Database with Aerospace;* arXiv*; Library & Information Science Collection; Computer and Information Systems Abstracts; Directory of Open Access Journals* – DOAJ; Elsevier – CrossRef; *Engineering Research Database; Materials Science & Engineering Database; Mechanical & Transportation Engineering Abstracts;* MEDLINE/PubMed – NLM; OneFile – GALE; PMC – PubMed Central; *Science Citation Index Expanded* – Web of Science; Scopus – Elsevier; *Social Sciences Citation Index* – Web of Science, Springer – CrossRef; SpringerLink; *Technology Research Database; Wiley Online Library*.					
Portal Ebsco com as bases *Academic Search Premier; Computers & Applied Sciences Complete; Information Science & Technology Abstracts* – ISTA; *Library, Information Science & Technology Abstracts with Full Text.*					

FONTE: A autora (2019)

Na primeira etapa, objetivou-se verificar o total de publicações orientadas à classificação de textos por meio de aprendizado de máquina. Desse modo, as

expressões de busca contemplaram os seguintes termos: “aprendizado de máquina”, “aprendizado automático”, “sistemas inteligentes”, “inteligência computacional", “classificação de texto”, “categorização de texto”, “mineração de texto”, “processamento de texto” e “processamento de linguagem natural”.

Na segunda etapa, e com vistas a focar as buscas em serviços públicos, foram adicionados os termos “serviços públicos”, “serviços municipais”, “serviços emergenciais", "serviços ao cidadão", "serviços governamentais e “atendimento ao cidadão”.

Em seguida, com o objetivo de recuperar teses e dissertações, foi realizada busca na base de dados e portal brasileiros: Biblioteca Digital Brasileira de Teses e Dissertações, do Instituto Brasileiro de Informação em Ciência e Tecnologia – BDTD/IBCTI e Catálogo de Teses e Dissertações, da Capes. As expressões e os termos utilizados são apresentados na Tabela 3.

TABELA 3 – EXPRESSÕES DE BUSCA, BASE DE DADOS E PORTAL BRASILEIROS UTILIZADOS NA PESQUISA DE TESES E DISSERTAÇÕES

Expressão de busca	Base de dados e portal brasileiros		
	BDTD/IBCTI	Capes	Total
(“aprendiza* de máquina” OR “aprendizado automático” OR “sistema* inteligente*” OR "inteligência computacional") AND (“classificação de texto*” OR “categorização de texto*” OR “mineração de texto*” OR “processamento de texto*” OR “processamento de linguagem natural”)	115	409	524
Busca focada			
AND (“serviços públicos” OR “serviços municipais” OR “serviços emergenciais" OR "serviços ao cidadão" OR "serviços governamentais OR “atendimento ao cidadão”)	0	21	21

FONTE: A autora (2019)

Em abril de 2021 houve uma tentativa de atualização desses resultados, contudo as bases *Advanced Technologies & Aerospace Database, Advanced Technologies Database with Aerospace, Library & Information Science Collection, Computer and Information Systems Abstracts,* Elsevier – CrossRef, *Engineering Research Database, Materials Science & Engineering Database, Mechanical & Transportation Engineering Abstracts,* Springer – CrossRef e *Technology Research Database* não estavam disponíveis no Portal de Periódicos da Capes, acesso via Comunidade Acadêmica Federada (CAFe), portanto, não foi possível repetir as buscas.

A revisão da literatura desta pesquisa pautou-se nos objetivos geral e específicos, nos conceitos e autores, de modo a auxiliar os encaminhamentos metodológicos, assjm, se faz relevante demonstrar esse alinhamento, conforme a síntese do Quadro 3.

QUADRO 3 – SÍNTESE DO ALINHAMENTO ENTRE OBJETIVOS ESPECÍFICOS, CONCEITOS E AUTORES

Objetivos	Conceitos	Autores
a) Descrever o processo de GI do atendimento ao cidadão, da Central 156	Informação, conhecimento, GI e modelo de GI	Choo (2003); Choo (2020); Davenport e Prusak (1998); Detlor (2010); Gallo (2010); Feldman e Sanger (2007); Marchiori (2002); McGee e Prusak (1994); Ponjuán Dante (2004); Setzer (1999); Silva e Ribeiro (2009); Sordi (2008).
	Relacionamento governo-cidadão por meio da tecnologia da informação	Agune e Carlos (2005); Diirr, Araujo e Capelli (2011); Fang (2002); Heringer *et al.* (2017); Ma e Zheng (2019); Meijer, Curtin e Hillebrandt (2012); Mendes Junior (2018); Mezzaroba e Bier (2016); Nam (2011); Prado *et al*. (2011); Ringold *et al.*(2012); Santos e Rover (2018); Schellong (2005); Schmidthuber *et al.* (2017); Sun, Ku e Shih (2015); Tavana, Zandi e Katehakis (2013); Vigoda (2002); Wu (2017).
b) Submeter os textos das demandas ao processamento de linguagem natural, representando-os no modelo espaço vetorial	PLN, níveis de entendimento e mineração de textos	Aggarwal e Zhai (2012); Androutsopoulou *et al.* (2019); Arampatzis *et al.* (2000); Aranha e Passos (2006); Baghdadi *et al.* (2019); Berti (2017); Chowdhury (2003); Fedman e Sanger (2007); Gonzalez e Lima (2003); Jackson e Moulinier (2002); Kano, Fujita e Tsuda (2019); Liddy (2001); Liddy (2002); Loh (2014); Manning e Schütze (1999); Monteiro (2017); Pollettini (2016); Weiss, Indurkya e Zhang (2015); Witten *et al.* (2017).
	Representação de documentos, *n*-gramas, ponderação de termos e redução de dimensionalidade	Aggarwal e Zhai (2012); Arampatzis *et al.* (2000); Awad e Khanna (2015); Baghdadi *et al.* (2019); Berti (2017); Feldman e Sanger (2007); Gonzalez e Lima (2003); Han, Karypis e Kumar (1999); Jackson e Moulinier (2002); Kano, Fujita e Tsuda (2019); Manning, Raghavan e Schütze (2009); Manning e Schütze (1999); Martins, Monard e Matsubara (2003); Monard e Baranauskas (2003); Monteiro (2017); Rossi (2015); Pollettini (2016); Russell e Norvig (2013); Santos (2019); Sebastiani (2002); Silva e Souza (2002); Tan, Wang e Lee (2002); Weiss, Indurkya e Zhang (2015); Witten (2004).
c) Aplicar algoritmos de aprendizado de máquina para classificação das demandas por órgão	Aprendizado de máquina e conhecimento	Castro e Ferrari (2016); Davenport e Pruzak (1998); Lopes, Santos e Pinheiro (2014); McCarthy e Feigenbaum (1990); Mitchell (1997); Monard e Baranauskas (2003); Pollettini (2016); Russell e Norvig (2013); Sordi (2008); Witten *et al.* (2017).
	Abordagens do aprendizado de máquina; aprendizado supervisionado e concepção de modelos	Androutsopoulou *et al.* (2019); Aranha (2007); Awad e Khanna (2015); Bramer (2016); Castro e Ferrari (2016); Feldman e Sanger (2007); Goldschmidt e Passos (2005); Larose e Larose (2014); Lopes, Santos e Pinheiro (2014); Manning, Raghavan e Schütze (2009); Monard e Baranauskas (2003); Prati (2006); Russell e Norvig (2013); Vasques *et al.* (2017).
	Classificação de textos, algoritmos e tipificação do conjunto de classes	Aggarwal e Zhai (2012); Awad e Khanna (2015); Baghdadi *et al.* (2019); Berti (2017); Bramer (2016); Castro e Ferrari (2016); Eyheramendy, Lewis e Madigan (2003); Feldman e Sanger (2007); Goldschmidt e Passos (2005); Han, Karypis e Kumar (1999); Larose e Larose (2014); Lima (2010); McCallum e Nigam (1998); Manning, Raghavan e Schütze (2009); Mitchell (1997); Monard e Baranauskas (2003); Pollettini (2016); Russell e Norvig (2013); Schneider (2005); Sebastiani (2002); Weiss, Indurkya e Zhang (2015); Wang (2006); Witten *et al.* (2017).

FONTE: A autora (2021)

3.2.2 Pesquisa documental

A pesquisa documental intencionou subsidiar o alcance do objetivo a) da dissertação. Para tanto, investigou-se como ocorre o processo de gestão da informação do atendimento ao cidadão da Central 156, seguindo-se o modelo proposto por McGee e Prusak (1994, p. 107-126). A pesquisa foi realizada mediante utilização das principais ferramentas do Sistema Integrado de Atendimento ao Cidadão-156 (SIAC-156), desenvolvido pelo Instituto das Cidades Inteligentes (ICI), bem como leitura do Manual de instrução do cadastrador (ICI, 2019a), do Manual de instrução do responsável pelo serviço no órgão (ICI, 2019b) e da Tabela de serviços (ICI, 2021a).

O SIAC-156 compreende um dos sistemas corporativos da prefeitura, sendo utilizado internamente por todos os órgãos para visualização dos protocolos com as demandas dos cidadãos, acompanhamento do trâmite dos protocolos e resposta para o cidadão. Outro uso compreende a geração de relatórios e mapas estratégicos que possibilitam análises em auxílio ao monitoramento e desenvolvimento de ações e programas da PMC.

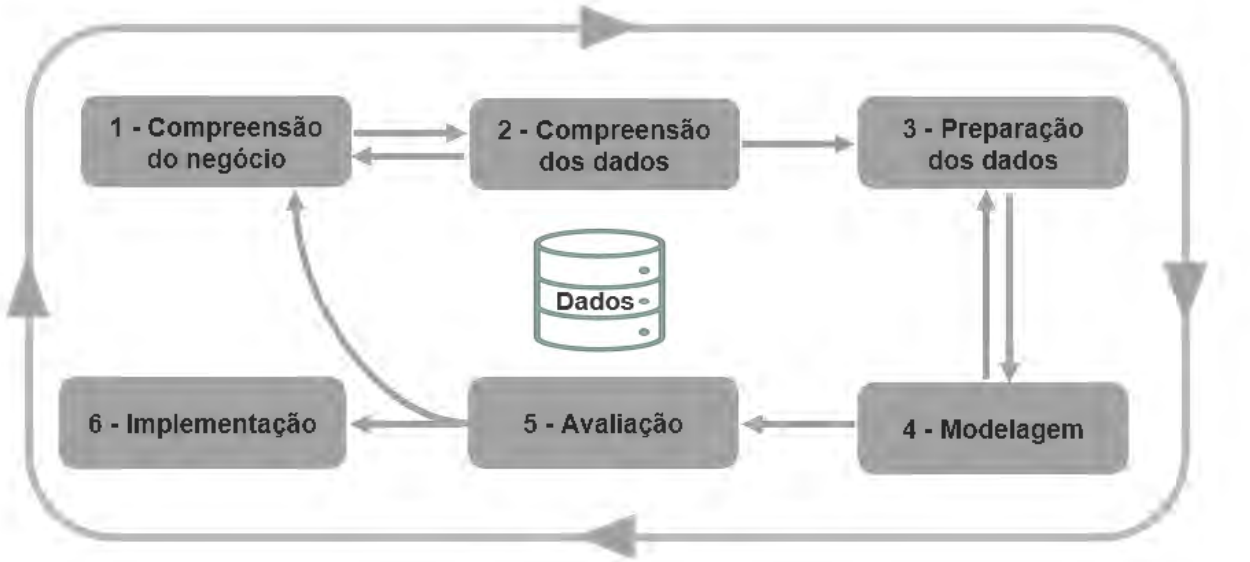
Na resposta para o cidadão pode constar o parecer sobre a solicitação, a previsão de realização do serviço ou a justificativa para a não realização. Há ainda a resposta administrativa, utilizada quando a natureza do serviço não envolve uma execução ou quando é referente a elogios ou reclamações.

3.2.3 Pesquisa experimental

A metodologia experimental deste trabalho é orientada pelo *Cross Industry Standard Process Data Mining* (CRISP-DM) ou Processo Padrão de Vários Segmentos de Mercados para Mineração de Dados, no português, e procura auxiliar o alcance dos objetivos b) e c) da dissertação.

O CRISP-DM constitui seis fases para desenvolvimento de projetos de mineração de dados e é flexível, o que permite fácil adaptação para trabalhos envolvendo aprendizado de máquina, PLN e mineração de textos. A Figura 11 apresenta as fases do CRISP-DM (IBM, 2015, p. 1) que são descritas a seguir.

FIGURA 11 – FASES DA METODOLOGIA CRISP-DM



FONTE: A autora (2020) adaptado de IBM (2015, p. 1)

Fase 1 – compreensão do negócio

Abrange o entendimento do negócio da organização, a identificação do(s) problema(s), a finalidade do trabalho, recursos disponíveis e os objetivos esperados, de modo a subsidiar a elaboração de um plano de trabalho (IBM, 2015, p. 5-6), correspondendo a um cronograma. O plano de trabalho deste estudo não foi apresentado, pois foi elaborado e seguido até a finalização da dissertação. O detalhamento desta fase consta no tópico 4.1.

Fase 2 – compreensão dos dados

Compreende a coleta, exploração e análise da qualidade dos dados (IBM, 2015, p. 13). A descrição desta fase é apresentada no tópico 4.2.

Fase 3 – preparação dos dados

Geralmente compreende a fase mais demorada do trabalho e, quando bem desenvolvida, evita gastos adicionais. Envolve as etapas de agrupar conjuntos de dados, selecionar subconjuntos ou amostras dos dados, agregar e derivar atributos, remover atributos ruidosos e dividir o conjunto de dados em treinamento e teste (IBM, 2015, p. 19). Na presente pesquisa a preparação dos dados engloba as etapas de pré-processamento dos textos em linguagem natural e a representação dos documentos no modelo espaço vetorial, detalhadas no tópico 4.3.

Fase 4 – modelagem

Trata da submissão do conjunto de dados às ferramentas de análise. Nessa fase são selecionados os algoritmos e gerados os modelos, necessitando de ajustes de parâmetros, o que leva a serem realizadas diversas iterações (IBM, 2015, p. 19). Nesta fase ocorreu o processamento dos algoritmos de aprendizado de máquina, conforme descrito no tópico 4.4.

Fase 5 – avaliação

Consiste em avaliar se os resultados obtidos atendem os objetivos estabelecidos na primeira fase do trabalho. Em caso negativo, retorna-se à primeira fase. Na avaliação são identificadas falhas, dificuldades e decisões ou estratégias alternativas que poderiam ter sido usadas (IBM, 2015, p. 31-32). Esta fase é detalhada no tópico 4.5.

Fase 6 – implementação

Os resultados são apresentados à organização e o modelo obtido poderá ser integrado aos sistemas de informações ou auxiliar no planejamento e na tomada de decisão. Por vezes, a dinâmica do negócio da organização exigirá atualização dos resultados, tornando o processo de avaliação do modelo constante.

Como trata-se de uma pesquisa acadêmica, posteriormente à sua conclusão, será realizada reunião com a Secretaria do Governo Municipal, o Instituto de Pesquisa e Planejamento Urbano de Curitiba e o Instituto das Cidades Inteligentes, visando apresentar os objetivos da pesquisa, o embasamento teórico, o desenvolvimento do método para elaboração do modelo proposto, assim como os resultados alcançados. A dissertação, tanto no formato impresso quanto digital, também será disponibilizada à PMC.

4 DESENVOLVIMENTO DO MÉTODO CRISP-DM

Este capítulo descreve o desenvolvimento do método para obtenção do modelo de classificação automática das demandas da Central 156. O capítulo estrutura-se conforme as fases do método CRISP-DM, com exceção da Fase 6 – implementação, descrita anteriormente no tópico 3.2.3.

4.1 COMPREENSÃO DO NEGÓCIO

Com vistas a possibilitar melhor entendimento do negócio, esta fase apresenta o ambiente da pesquisa, os canais disponibilizados pela Central 156 para relacionamento com os cidadãos, os fluxos de atendimento existentes e a distribuição das demandas entre os canais. Também são elencados os recursos disponíveis para desenvolvimento do método.

4.1.1 Ambiente da pesquisa

O ambiente da pesquisa abrange a Central 156 de Atendimento ao Cidadão da Prefeitura Municipal de Curitiba. Com vistas possibilitar melhor compreensão e contextualizar esse ambiente, são apresentadas algumas informações acerca da cidade e da prefeitura.

Curitiba é a capital do Estado do Paraná e localiza-se na Região Sul do Brasil, a 1.400 km de Brasília, a capital do país, e 410 km de São Paulo. Fundada em 1693, possui 435 km^2 e população próxima dos dois milhões de habitantes (PMC, 2021a). A Prefeitura de Curitiba abrange 29 órgãos entre secretarias, autarquias, fundações, sociedades de economia mista e os gabinetes do prefeito e do vice e, ainda, 10 administrações regionais (PMC, 2021b). As administrações regionais estão distribuídas estrategicamente no território, de modo a ofertar serviços descentralizados aos cidadãos, tais como de infraestrutura urbana, ambiental, econômica e social. Também fazem parte da estrutura da prefeitura cerca de 2.300 equipamentos municipais, com unidades de atendimento das áreas saúde, educação, cultura, assistência social e esporte e lazer (IPPUC, 2021).

Em dezembro de 2020, trabalhavam na PMC 29.690 servidores (PMC, 2020), sendo a receita orçamentária prevista, para o exercício de 2021, de R$ 8,127 bilhões (Curitiba, 2020, p. 2).

Em Curitiba, com vistas a estreitar e facilitar o relacionamento entre prefeitura e cidadãos, a prestação dos serviços públicos, além de presencial, é realizada via telefone, e-mail, portal, aplicativo, rede social e pela Central 156.

A Central 156, ambiente deste estudo, foi criada em 1984, como uma subdivisão do serviço denominado Tele Documentos, sob responsabilidade do Departamento de Fazenda, hoje Secretaria Municipal de Planejamento, Finanças e Orçamento. Por telefone, o Tele Documentos fornecia informações sobre alvarás de funcionamento e, pelo correio, encaminhava certidões negativas de débitos fiscais, solicitadas pelos contribuintes de Curitiba (SGM, 2015, p. 1).

O projeto do novo modelo de atendimento foi desenvolvido pelo antigo Centro de Processamento de Dados (CPD), à época vinculado ao Instituto de Pesquisa e Planejamento Urbano de Curitiba, para diminuir as filas no balcão de atendimento da prefeitura (IPPUC, 1991, p. 1). O objetivo era registrar solicitações e reclamações relacionadas aos serviços prestados pelos diversos órgãos da prefeitura, bem como prestar informações pertinentes ao âmbito da administração municipal. No início, uma linha telefônica convencional era utilizada, sendo substituída, posteriormente, pelo número 156, selecionado a partir de uma listagem de possíveis números encaminhados pela Telepar, antiga empresa operadora de telefonia no Paraná (SGM, 2015, p. 1).

Um único atendente, durante o horário comercial, recebia a ligação e anotava a solicitação ou reclamação, além do nome, endereço e telefone do cidadão e os monitores eram encarregados de “ir atrás” das respostas, que eram transmitidas por telefone ou carta (SGM, 2015, p. 2).

Posteriormente, os dados passaram a ser digitados em terminais IBM e enviados aos setores competentes, por meio de formulários para preenchimento das respostas. Esses formulários, após preenchidos, retornavam à Central para cadastramento, possibilitando emissão de relatórios gerenciais, acompanhamento e encaminhamento ao cidadão, via correio (IPPUC, 1991, p. 2).

Com o tempo, uma listagem dos assuntos solicitados pelos cidadãos foi criada pelo CPD e os primeiros assuntos inseridos foram reclamação de ônibus, coleta de

poda de árvores e lixo de jardim, iluminação pública e troca de lâmpada. Esses assuntos continuam entre os mais requisitados até hoje (SGM, 2015, p. 2).

Nos anos seguintes, mais atendentes e linhas telefônicas foram contratados, o atendimento foi ampliado para três turnos, operando das 8h às 23h (SGM, 2015, p. 2). Em 1999, a Central 156 foi transferida para o Instituto Curitiba de Informática, hoje Instituto das Cidades Inteligentes (ICI). Em 2002 passou a atender 24 horas por dia, sete dias por semana, com atendentes atuando em turnos de seis horas. Todos os órgãos da prefeitura são integrados à Central, ocorrendo o encaminhamento *on-line* das demandas aos responsáveis por analisar, fiscalizar, executar e elaborar os pareceres (SGM, 2015, p. 4).

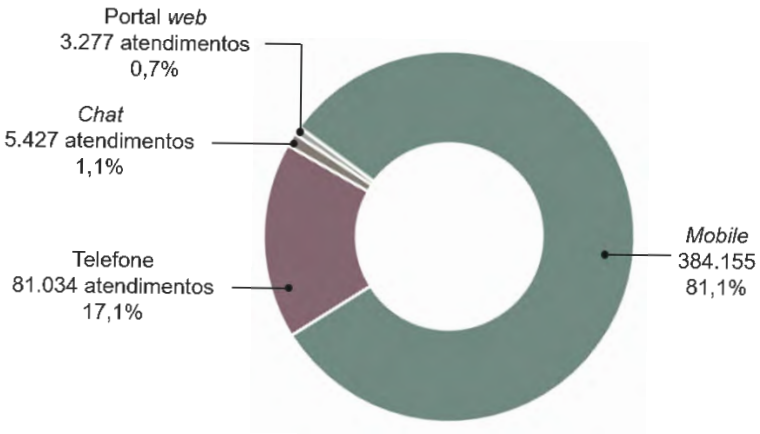
A Central 156 dispõe de cinco canais de atendimento ao cidadão, compreendendo o nível governo eletrônico de relacionamento governo-cidadão. Além do telefone 156, é possível registrar as demandas pelo *chat*, que apresenta duas modalidades, o humano e o automatizado, bem como pelo portal *web*, no endereço *http://www.central156.org.br/*, e pelo aplicativo *mobile* Curitiba 156, implantado em março de 2019. As demandas se dividem em informações e serviços, sendo as informações resolvidas, geralmente, durante o atendimento telefônico ou diretamente no canal *mobile*. Itinerários, horários das linhas de ônibus e localização dos ônibus, em tempo real, são exemplos de informações disponibilizadas no *mobile*. As demandas por serviços, que compreendem aos dados analisados neste estudo, em 2019, de acordo com os dados do Portal de Dados Abertos da PMC se subdividiam-se em solicitações, com 87% do total, reclamações 8,84%; elogios 2,44%; recadastros 1,31%; sugestões 0,28% e; denúncias 0,23% (PMC, 2019a).

Para visualizar a segmentação das demandas entre os canais, foram utilizados os números de atendimentos do relatório Estatísticas & Indicadores (ICI, 2019c), referente a dezembro de 2019. Esses atendimentos contemplam demandas por informações e serviços. O Gráfico 3 apresenta a segmentação entre os canais.

O fluxo de atendimento difere entre os canais. Por telefone e *chat* humano é possível requisitar qualquer serviço da lista e, nesses casos, as demandas são classificadas, isto é, rotuladas manualmente pelos atendentes. Conforme planilhas disponibilizadas no Portal de Dados Abertos (PMC, 2019a), para o ano de 2019 havia 1.128 serviços, contudo, o rol de serviços chega a 2.500 categorias. O *chat* automatizado permite demandas para três serviços: pavimentação; coleta de resíduos vegetais; e iluminação pública. Pelo portal *web* estão disponíveis 68 serviços, não

sendo possível consultar informações. No *mobile* são 44 os serviços disponíveis e há possibilidade de consultar informações como telefones úteis, locais de atendimento da PMC e horário e itinerário de ônibus.

GRÁFICO 3 – ATENDIMENTOS REALIZADOS PELA CENTRAL 156 DE CURITIBA, SEGUNDO CANAIS DE ATENDIMENTO, NO MÊS DE DEZEMBRO DE 2019



FONTE: A autora (2021) com base em ICI (2019c)

Nos atuais fluxos de atendimento da Central 156, demandas condizentes a reclamações, denúncias e elogios são abertas exclusivamente por telefone ou *chat* humano e registradas somente pelos atendentes, uma vez que requerem descrição detalhada de determinada situação. Nos canais digitais, somente solicitações podem ser registradas a partir da seleção de opções preestabelecidas. Os fluxos simplificados dos canais de atendimento da Central 156 encontram-se nos Apêndices 1, 2, 3 e 4. Esses fluxos constituem importantes produtos informacionais resultantes da pesquisa.

A gestão dos sistemas, da base de dados e dos *webservices* utilizados pela Central 156 são de responsabilidade do ICI. A gestão do atendimento, dos serviços existentes e das demandas é da PMC, cabendo à Secretaria Municipal do Governo Municipal a coordenação, definição de normativas e a articulação, entre os órgãos, do atendimento ao cidadão pela prefeitura.

Para identificação das tarefas que compõem o fluxo do processo de gestão da informação do atendimento ao cidadão, em consonância ao modelo de McGee e Prusak (1994, p. 107-126), foram exploradas as funcionalidades do SIAC-156 e seus manuais de utilização. O sistema apresenta menu de navegação para cadastro e consulta das demandas dos cidadãos, conforme demonstrado na Figura 12.

FIGURA 12 – FUNCIONALIDADES DO SIAC-156

FONTE: SIAC-156 (ICI, 2021b)

O menu “Cadastro” permite registrar:

- dados automáticos: código do protocolo, do cadastrador e data;
- dados do solicitante: pessoa física (nome, CPF, data de nascimento, endereço, telefone); pessoa jurídica (nome da empresa e nome fantasia, CNPJ, ramo de atividade, nome do contato na empresa, endereço, telefone).

Por meio do menu “Movimentação”, submenu “Solicitação”, são registrados os dados das demandas, conforme apresentado na Figura 13.

Os dados de cada demanda contemplam:

- tipo: que pode ser solicitação, informação, sugestão, denúncia, elogio, reclamação ou consulta pública;
- descrição: conteúdo, em formato texto, da solicitação, permitindo até 5.000 caracteres. Dependendo do serviço da demanda, o sistema apresenta textos com descrições padrões para solicitação, com vistas a agilizar o preenchimento desse dado, permitindo alteração conforme necessidade;
- assunto: assunto específico conforme a descrição do que está sendo solicitado;
- subdivisão: associada ao assunto, compreende um de seus subassuntos;
- local de atendimento: logradouro, número, CEP, transversais do endereço, bairro, ponto de referência. O cadastro é realizado somente quando divergir do endereço do solicitante.
- dados importantes: por exemplo, nome da escola ou da unidade de saúde;

- meio de resposta: telefone, e-mail ou pessoalmente;
- solicita sigilo: sim ou não;
- resposta: são apresentadas todas as respostas emitidas durante as tramitações do protocolo, da mais recente para a mais antiga;
- histórico: são apresentadas todas as movimentações realizadas a partir de sua criação. É possível visualizar os encaminhamentos, consultas e trâmites que foram executados no documento até o seu encerramento.

FIGURA 13 – TELA DE REGISTRO DA DEMANDA NO SIAC-156, COM DESTAQUE PARA O CAMPO DESCRIÇÃO

FONTE: SIAC-156 (ICI, 2021b)

O menu “Tabelas” permite acesso às tabelas auxiliares do sistema, como a de serviços, com busca orientada por assuntos, subdivisões, órgãos e palavras-chave. Dependendo do nível de acesso, são liberadas funcionalidades específicas para órgãos que demandam informações não utilizadas corporativamente, como é o caso das Secretarias de Educação e da Saúde.

4.1.2 Recursos utilizados para desenvolvimento do método CRISP-DM

Quanto aos recursos, o equipamento configura-se como um microcomputador laptop Intel (R) Core (TM) i3 – 1005G1, com CPU de 1.20 GHz, RAM de 8.00 Gb, sistema operacional 64b e Windows 10. As ferramentas incluem os softwares Access e Excel, do pacote Microsoft Office, o software estatístico de código aberto R, versão 3.6.2 e o software *Waikato Environment for Knowledge Analysis* ou Weka, versão 3.8.3, que compreende uma coleção de algoritmos de aprendizado de máquina para tarefas de mineração de dados.

4.2 COMPREENSÃO DOS DADOS

Esta fase descreve a coleta, exploração e análise dos dados, e os passos realizados para obtenção da amostra do estudo.

Os dados utilizados na pesquisa foram coletados a partir do Portal de Dados Abertos da PMC e correspondem a 2019, ano de início deste estudo. No portal, arquivos em formato .csv são disponibilizados mensalmente, tendo como referência a movimentação do mês anterior, contudo, em 2019 havia dois arquivos correspondendo ao mês de junho e nenhum ao mês de maio. Dessa maneira, a junção dos 11 arquivos .csv totalizou 316.299 registros.

Esses arquivos apresentam 21 campos, conforme dicionário de dados apresentado no Quadro 4 e também disponibilizado no Portal de Dados Abertos da PMC.

QUADRO 4 – DICIONÁRIO DOS DADOS ABERTOS DA CENTRAL 156 DE CURITIBA

Seq.	Campo	Descrição	Tipo (tamanho)
1	SOLICITACAO	Código da demanda	Integer
2	TIPO	Tipo da demanda (solicitação, informação, sugestão, elogio, denúncia, reclamação, consulta pública)	Varchar(40)
3	ORGAO	Órgão do cadastrador da demanda	Varchar (80)
4	DATA	Data de criação da demanda	Datetime
5	HORARIO	Hora de criação da demanda	Datetime
6	ASSUNTO	Assunto ao qual a demanda se refere	Varchar (100)
7	SUBDIVISÃO	Subdivisão do assunto ao qual a demanda se refere	Varchar (150)
8	DESCRICAO	Descrição da demanda	Text
9	LOGRADOURO_ASS	Logradouro de localização da demanda	Varchar (255)
10	BAIRRO_ASS	Bairro de localização da demanda	Varchar (50)
11	REGIONAL_ASS	Regional de localização da demanda	Varchar (50)
12	MEIO_RESPOSTA	Meio escolhido para encaminhamento da resposta ao cidadão (telefone, e-mail, pessoalmente)	Varchar (15)
13	OBSERVAÇÃO	Observações referentes à demanda	Text
14	SEXO	Sexo do cidadão (M – Masculino, F - Feminino)	Varchar (1)
15	BAIRRO_CIDADAO	Bairro do domicílio do cidadão	Varchar (50)
16	REGIONAL_CIDADAO	Regional do domicílio do cidadão	Varchar (50)
17	DATA_NASC	Data de nascimento do cidadão	Datetime
18	TIPO_CIDADAO	Tipo de solicitante (cidadão, vereador, polícia, hospital etc)	Varchar (30)
19	HISTORICO	Última situação da demanda (responder para o solicitante, envio do e-mail de resposta, conclusão automática etc)	Varchar (100)
20	ORGAO_RESP	Órgão responsável pela resposta	Varchar (80)
21	RESPOSTA_FINAL	Descrição da resposta	Text

FONTE: PMC (2019b)

Desses 21 campos, foram utilizados inicialmente seis: código do protocolo, órgão responsável, serviço, descrição da demanda, tipo de demanda e data. Os quatro primeiros campos foram escolhidos por apresentar os dados necessários para a classificação das demandas. O campo tipo de demanda por possibilitar a seleção apenas dos registros referentes à “solicitação”, uma vez que havia no material registros referentes ao tipo “informação”. A data para exclusão dos registros com ano diferente de 2019.

No Quadro 5 é apresentado um exemplo de demanda registrada pela Central 156 em 2019.

QUADRO 5 – EXEMPLO DE DEMANDA REGISTRADA PELA CENTRAL 156 EM 2019

Seq.	Campo	Dado
1	SOLICITACAO	7673999
2	TIPO	SOLICITAÇÃO
3	ORGAO	INSTITUTO DAS CIDADES INTELIGENTES
4	DATA	05/02/2019
5	HORARIO	09:01:18
6	ASSUNTO	PAVIMENTAÇÃO
7	SUBDIVISÃO	MANUTENÇÃO
8	DESCRICAO	SOLICITA MANUTENÇÃO DE PAVIMENTO. INFORMA QUE O BURACO ESTÁ LOGO APÓS UMA LOMBADA.
9	LOGRADOURO_ASS	REINALDO HECKE, 753
10	BAIRRO_ASS	ABRANCHES
11	REGIONAL_ASS	UNIDADE REGIONAL BOA VISTA
12	MEIO_RESPOSTA	E-MAIL
13	OBSERVAÇÃO	PRÓXIMO AO PONTO DE REFERÊNCIA
14	SEXO	F
15	BAIRRO_CIDADAO	BAIRRO ALTO
16	REGIONAL_CIDADAO	UNIDADE REGIONAL BOA VISTA
17	DATA_NASC	14/09/1984
18	TIPO_CIDADAO	CIDADÃO
19	HISTORICO	CONCLUSÃO AUTOMÁTICA
20	ORGAO_RESP	SECRETARIA DO GOVERNO MUNICIPAL
21	RESPOSTA_FINAL	SOLICITAÇÃO ATENDIDA

FONTE: PMC (2019a)

Para obtenção da amostra do estudo, preliminarmente, houve eliminação dos registros inconsistentes ou não relevantes, considerando-se: registros replicados, com quebras de linha indevidas, com tipo de demanda igual a “consulta pública” ou “informação”, do ano de 2020 e com o campo descrição preenchido como “sigiloso”, contemplando apenas números ou com menos de 10 caracteres. A Tabela 4 apresenta o total de registros eliminados em cada uma dessas etapas.

Devido à associação incorreta de um serviço a mais de um órgão, isto é, registros com demandas iguais classificadas em órgãos diferentes, inclusive violando a integridade referencial entre as tabelas, foi realizada correção manual dos registros nessa condição. Nesta etapa não houve eliminação de registros.

TABELA 4 – ELIMINAÇÃO DE REGISTROS INCONSISTENTES OU NÃO RELEVANTES

Situação	Total
Total inicial de registros (junção dos 11 arquivos de 2019 em formato .csv)	316.299
Exclusão de registros	
Replicados	34.139
Gerados a partir de quebras de linha indevidas	5.154
Com o campo tipo de demanda = “consulta pública”	4
Com o campo tipo de demanda = “informação”	1.018
Do ano de 2020	7
Com o campo descrição = “sigiloso”, números e menos de 10 caracteres	1.984
Total	273.993

FONTE: A autora (2021)

Em seguida foi analisada a distribuição das demandas com o propósito de classificá-las por órgão. Em virtude da fusão de órgãos ao longo do ano de 2019, havia diferentes nomenclaturas para as mesmas atribuições no organograma da PMC. Desse modo, foram agrupadas as demandas dos órgãos afins, totalizando 28 órgãos, dos 34 iniciais, e 273.993 demandas, conforme demonstrado na Tabela 5.

Apenas oito órgãos são responsáveis por 98% do total das demandas, evidenciando que as classes são bastante desbalanceadas, implicando num conjunto de dados com classes majoritárias e minoritárias. Desse modo, para efetuar a classificação, foram considerados os oito órgãos com maiores demandas, correspondendo a 268.485 exemplos. Dentre essas oito classes, também é possível observar um grande desbalanceamento entre o número de exemplos. A diferença significativa entre o número de exemplos de cada classe não é desejável quando as classes minoritárias possuem informações relevantes (MONARD; BARANAUSKAS, 2003, p. 46). Nesses casos, os algoritmos de aprendizado podem encontrar dificuldades para classificar exemplos das classes minoritárias, criando modelos de baixo desempenho, que quase sempre predizem a classe majoritária (PRATI, 2006, p. 87-88).

Para atingir o equilíbrio entre as classes, foi gerada uma amostra com *undersampling* tendo como base o número de exemplos da classe minoritária, que corresponde à Secretaria Municipal do Urbanismo.

TABELA 5 – DISTRIBUIÇÃO DAS DEMANDAS DO 156 SEGUNDO ÓRGÃOS DA PMC EM 2019

Seq.	Órgãos	Demandas	%
1	Secretaria Municipal do Meio Ambiente (SMMA)	76.175	27,80
2	Secretaria Municipal da Defesa Social e Trânsito (SMDT)*	60.568	22,11
3	Secretaria Municipal de Obras Públicas (SMOP)	38.589	14,08
4	Fundação de Ação Social (FAS)	27.041	9,87
5	Secretaria Municipal da Saúde (SMS)	26.081	9,52
6	Secretaria do Governo Municipal (SGM)	17.482	6,38
7	Urbanização de Curitiba (URBS)	11.668	4,26
8	Secretaria Municipal do Urbanismo (SMU)	10.881	3,97
9	Secretaria Municipal da Educação	1.697	0,62
10	Órgãos de Administração**	1.355	0,49
11	Secretaria Municipal do Esporte, Lazer e Juventude	808	0,30
12	Secretaria Municipal de Segurança Alimentar e Nutricional***	742	0,27
13	Secretaria Municipal do Trabalho e Emprego	262	0,10
14	Companhia de Habitação Popular de Curitiba	151	0,06
15	Instituto de Pesquisa e Planejamento Urbano de Curitiba	115	0,05
16	Instituto Municipal de Turismo - Curitiba turismo	91	0,03
17	Fundação Cultural de Curitiba	60	0,02
18	Instituto Curitiba de Saúde	56	0,02
19	Instituto de Previdência dos Servidores do Município de Curitiba	46	0,02
20	Secretaria Municipal da Comunicação Social	34	0,01
21	Procuradoria Geral do Município	26	0,01
22	Secretaria Especial dos Direitos da Pessoa com Deficiência	25	0,01
23	Instituto Municipal de Administração Pública	13	0,00
24	Instituto das Cidades Inteligentes****	10	0,00
25	Agência Curitiba de Desenvolvimento S/A	8	0,00
26	Secretaria Municipal de Recursos Humanos	5	0,00
27	Secretaria de Informação e Tecnologia	3	0,00
28	Companhia de Desenvolvimento de Curitiba	1	0,00
Total		273.993	100,00

* Compreende demandas da Secretaria Municipal da Defesa Social, Secretaria Municipal de Trânsito e Secretaria Municipal da Defesa Social e Trânsito.

** Classe estabelecida pela autora. Compreende demandas da Secretaria Municipal de Administração e Gestão de Pessoal, Secretaria Municipal de Finanças, Secretaria Municipal de Planejamento e Administração e Secretaria Municipal de Planejamento, Finanças e Orçamento.

*** Compreende demandas da Secretaria Municipal do Abastecimento e Secretaria Municipal de Segurança Alimentar e Nutricional.

**** Associação civil sem fins lucrativos que atua em parceria com a PMC por meio de contratos de gestão (ICI, 2021c).

FONTE: A autora (2021) com base em PMC (2019a).

Para eliminação dos exemplos, foram consideradas as quantidades de exemplos existentes em cada serviço de cada uma das oito classes. Lembrando que os serviços são formados pela associação dos assuntos e das subdivisões de cada órgão. Com fundamento em Feldman e Sanger (2007, p. 70), os serviços com menos de 36 exemplos foram desprezados. Os autores apontam que, como regra geral, 30 exemplos são necessários em cada classe para treinamento do modelo. Dessa maneira, a classe minoritária, que tinha 10.881 exemplos, passou a totalizar 10.547 exemplos.

Permanecendo os serviços com 36 ou mais exemplos e considerando-se uma amostra com 95% de grau de confiança e 5% de margem de erro, foram obtidas as quantidades mínimas de exemplos em cada serviço. Como a soma dessas quantidades ainda era muito maior que a quantidade de exemplos da classe minoritária, buscou-se identificar, dentre os serviços majoritários de cada classe, uma quantidade de exemplos que, multiplicada pelo número desses serviços majoritários e somada ao número de exemplos dos serviços minoritários, se aproximasse do valor da classe minoritária, isto é, 10.547 exemplos. A distribuição dos exemplos entre as classes, após a amostragem *undersampling,* totalizou 86.345 registros e pode ser visualizada na Tabela 6. No Apêndice 5 são expostos os cálculos de *undersampling* para a classe Secretaria Municipal do Meio Ambiente e no Apêndice 6 podem ser visualizados os cálculos para as demais classes.

TABELA 6 – DISTRIBUIÇÃO DOS EXEMPLOS ENTRE AS OITO CLASSES

Seq	Classe (órgão)	Inicial para as oito classes		Amostragem *undersampling*		Amostragem aleatória	
		Total	%	Total	%	Total	%
1	SMMA	76.175	28,37	11.921	13,81	5.155	13,72
2	SMDT	60.568	22,56	10.464	12,12	4.659	12,40
3	SMOP	38.589	14,37	11.022	12,76	4.778	12,71
4	FAS	27.041	10,07	9.306	10,78	4.193	11,16
5	SMS	26.081	9,71	11.767	13,63	5.061	13,46
6	SGM	17.482	6,51	10.765	12,47	4.598	12,23
7	URBS	11.668	4,35	10.553	12,22	4.591	12,21
8	SMU	10.881	4,06	10.547	12,21	4.553	12,11
Total		268.485	100,00	86.345	100,00	37.588	100,00

FONTE: A autora (2021)

Devido à limitação computacional, posteriormente foi gerada uma amostra aleatória a partir da amostra completa gerada na etapa anterior. Nessa amostragem foi utilizado 99% de grau de confiança e 0,5% de margem de erro, totalizando 37.588 exemplos. Os 37.588 exemplos compreendem a amostra utilizada nesta pesquisa.

4.3 PREPARAÇÃO DOS DADOS

A partir da amostra do estudo, com as descrições das demandas, nesta fase realiza-se o pré-processamento dos textos em linguagem natural e a representação dos documentos no modelo espaço vetorial.

Tendo em vista a transformação do conjunto de demandas numa representação processável pelos algoritmos de classificação, num primeiro momento, os textos foram submetidos ao pré-processamento. Nessa etapa, geralmente, ocorre a limpeza dos textos por meio da execução de uma série de funções vinculadas à mineração de textos descritas no tópico 2.6.

Como esta pesquisa utiliza o nível de entendimento morfológico do processamento de linguagem natural, ainda no Access, foram removidos dos textos plurais e formas femininas, indicados entre parênteses, e pronomes oblíquos, tanto no singular quanto no plural. E-mails e links para páginas da internet também foram removidos.

Devido à disponibilidade de funções para tratamento de textos, na continuidade do pré-processamento, optou-se pela utilização do software R. Com adição do pacote tm (*text mining*) ao R, o conjunto de textos das demandas foi convertido em c*orpus.* O pacote tm possibilita organizar, transformar e analisar dados textuais, fornecendo a integração necessária aos principais métodos estatísticos do R (FEINERER *et al.*, 2008, p. 4). Dessa maneira, foram aplicadas as seguintes funções de limpeza:

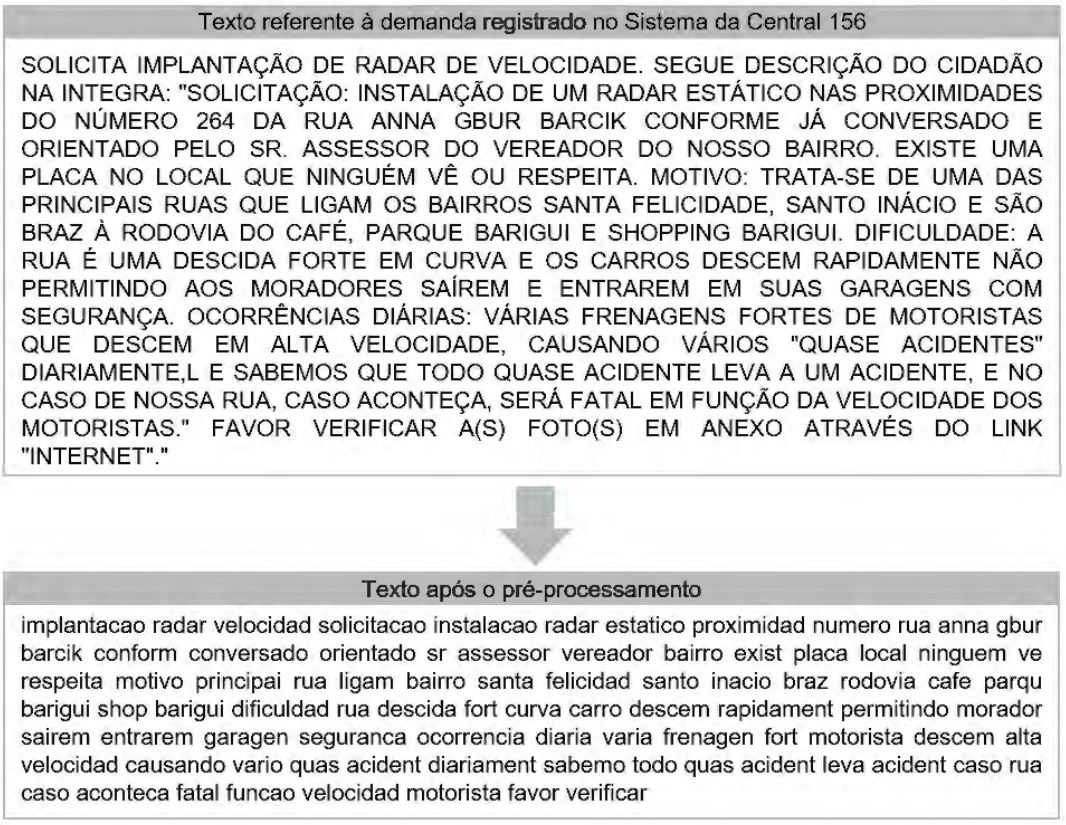
- alteração dos caracteres para minúsculo;
- remoção de caracteres especiais e pontuação;
- remoção de números;
- substituição dos caracteres acentuados do idioma português;
- remoção das *stopwords* a partir da lista em português (Apêndice 7), na qual foram adicionadas palavras do domínio do 156 com pouca

significância e grande quantidade de repetições como “cidadão”, “descrição” e “solicitação”;

- *stemming*[7], para igualar as variantes morfológicas, reduzindo as palavras ao seu radical, e remoção de espaços a mais entre as palavras dos documentos.

O resultado da limpeza é exemplificado na Figura 14.

FIGURA 14 – EXEMPLIFICAÇÃO DO PRÉ-PROCESSAMENTO DO TEXTO DA DEMANDA DO 156



FONTE: A autora (2021)

Após a limpeza e permanecendo o conteúdo mais significativo do *corpus*, este foi transformado na representação estruturada do espaço vetorial, ou matriz de termos e documentos, conforme modelo apresentado no tópico 2.5. Uma abordagem bastante utilizada para criar a matriz é *tokenization*, ou tokenização no português. Na tokenização, o texto é quebrado em *tokens* ou termos que constituem-se por palavras simples, compostas, sentenças e outros símbolos. Neste estudo, a representação foi

[7] A função stemDocument() do pacote tm do software R, utiliza o algoritmo de *stemming* de Porter (FEINERER; HORNIK, 2019, p. 38).

gerada com palavras simples - unigramas e com duas palavras compostas - bigramas. Na utilização das funcionalidades descritas, foi realizado tratamento para o idioma português no pacote tm e no RWeka.

Com a finalidade de distinguir os termos de menor ocorrência, porém, com maior significância, na geração da matriz de termos e documentos foi empregada a ponderação TF-IDF, com a qual são atribuídos pesos aos termos, sendo:

$$TF = tf_{i,j} \tag{2}$$

$$IDF = \log_2 \left(\frac{|D|}{|\{d|t_i \in d\}|} \right) \tag{3}$$

Onde:

$tf_{i,j} =$ número de ocorrências do termo t_i no documento d_j;

$|D| =$ total de documentos;

$|\{d|t_i \in d\}| =$ número de documentos onde o termo t_i aparece.

TF-IDF é calculada pelo produto da frequência do termo TF à frequência inversa no documento IDF (FEINERER; HORNIK, 2019, p. 56) e apresenta: a) valor mais alto quando o termo ocorre muitas vezes num número pequeno de documentos, dando alto poder discriminatório a esses documentos; b) valor baixo quando o termo ocorre poucas vezes em um documento e; c) valor baixíssimo quando o termo ocorre em quase todos os documentos (MANNING; RAGHAVAN; SCHÜTZE, 2009, p. 118-119).

A maioria dos atributos de uma matriz de termos e documentos é preenchida com zeros, significando que a maioria dos termos aparece em pouquíssimos documentos. Com muitos termos, e a maioria esparsos, a matriz apresenta alta dimensionalidade, o que tende a tornar o processamento computacional custoso e, por vezes, impraticável. Geralmente a taxa de esparsidade de uma matriz de termos e documentos é acima de 95%.

Neste estudo, como meios para reduzir a dimensionalidade, além da remoção de *stopwords* e conflação por *stemming,* procurou-se eliminar termos esparsos. Nesse caso, a questão é saber quantos desses termos devem ser eliminados, visto que

termos raros podem apresentar valor discriminatório. A função do pacote tm para essa finalidade necessita de um valor percentual limite para remoção dos termos, maior que zero e menor que um, de modo que, quanto mais próximo de um, menos termos serão eliminados. Por exemplo, quando fornecido como parâmetro o percentual 0,975, apenas os termos com esparsidade maior que este valor são removidos, isto é, a maioria dos termos esparsos será mantida. Por outro lado, se aplicado o valor 0,33, apenas os termos que ocorrem em dois terços dos documentos são mantidos. O cálculo de dispersão é dado pelo quociente do total de termos esparsos à dimensão da matriz. Sob essa perspectiva, foram realizadas tentativas para identificar os valores limites mais adequados, observando-se o número de termos resultantes e a capacidade computacional existente.

Assim sendo, na função removeSparseTerms() do R, foram utilizados dois limites de esparsidade, tanto para unigramas quanto para bigramas. Os valores limites utilizados para unigramas foram 0,999 e 0,9975, e para bigramas 0,999 e 0,9995. Dessa maneira, a partir da conversão das DTMs, foram gerados quatro *datasets* para o processamento dos algoritmos.

4.4 MODELAGEM

Nesta fase, os dados presentes nos *datasets,* gerados na fase anterior, foram submetidos à aplicação dos algoritmos de aprendizado, com o propósito de classificar as demandas automaticamente. Nesse sentido, esta pesquisa utiliza o aprendizado de máquina supervisionado, a inferência de aprendizado indutiva e os paradigmas de aprendizado simbólico, baseado em instâncias e o estatístico.

Optou-se pela utilização do software Weka, desenvolvido na Universidade de Waikato, Nova Zelândia (WITTEN *et al.* 2017, p. xxiv), visto que contempla uma diversidade de algoritmos. O processamento ocorreu após exportação das DTMs no R para o formato .csv e carregamento no Weka, a partir desse mesmo formato.

A técnica utilizada para treinamento e teste dos classificadores foi a validação cruzada. Foram utilizadas 10 partições, fazendo com que o processo fosse executado 10 vezes, considerando, a cada vez, nove partições para treinamento e uma para teste.

Os experimentos foram realizados com:

- o algoritmo baseado em árvores de decisão J48, uma vez que algoritmos dessa natureza são geralmente os primeiros a serem utilizados em experimentos de classificação (RUSSELL; NORVIG, 2013, p. 821);
- o algoritmo baseado em instâncias IBk (*Instance Based k-Nearest Neighbor*), visto que o k-NN é um dos métodos mais relevantes na classificação de textos (WEISS; INDURKHYA; ZHANG, 2015, p. 45), sendo amplamente utilizado (CASTRO; FERRARI, 2016, p. 168) e bastante efetivo, embora necessite de grande potencial computacional (WEISS; INDURKHYA; ZHANG (2015, p. 53);
- os algoritmos bayesianos Naïve Bayes e Naïve Bayes Multinomial, visto que são frequentemente utilizados em experimentos para classificação de textos, devido à sua simplicidade e eficácia (SCHNEIDER, 2005, p. 682). Optou-se pela utilização e análise do Naïve Bayes Multinomial, pois, nos experimentos iniciais, apresentou melhor desempenho que o Naïve Bayes. Naïve Bayes Multinomial é uma versão incremental do algoritmo baseado no Teorema de Bayes que, além da presença ou ausência, considera a frequência da palavra ou do termo no documento, haja vista que a frequência é uma informação potencialmente útil para determinar a classe de um documento (WITTEN *et al.,* 2017, p. 103).

4.5 AVALIAÇÃO

Como a tarefa de classificação automática de textos é, tipicamente, conduzida de modo experimental, métricas de desempenho baseadas em estatística se fazem necessárias para validar o resultado dos classificadores. Sebastiani (2002, p. 32) afirma que a avaliação experimental de um classificador geralmente mede sua efetividade, isto é, sua capacidade de decidir corretamente a classificação.

A matriz de confusão oferece uma medida efetiva do modelo de classificação ao demonstrar o número de classificações corretas em comparação ao número de classificações preditas, para cada classe do conjunto de exemplos (MONARD; BARANAUSKAS, 2003, p. 47). Nessa matriz, os valores da linha diagonal principal equivalem às predições corretas efetuadas pelo classificador e, nas demais células, encontram-se os valores das classificações incorretas. A partir da matriz de confusão

são derivadas métricas de avaliação, dentre essas a taxa de acerto do classificador e o coeficiente de Kappa, as quais foram utilizadas neste trabalho.

A taxa de acerto, ou acurácia, corresponde ao percentual de exemplos classificados corretamente. É obtida pela divisão do total de exemplos classificados corretamente, pelo número total de exemplos.

O coeficiente de Kappa foi apresentado por Cohen em 1960 e mede a concordância entre as classificações preditas e observadas, corrigindo a concordância que ocorre ao acaso (WITTEN *et al.,* 2017, p. 181). Em outras palavras, o coeficiente retira da taxa de acerto a probabilidade de classificação concordante ao acaso. É utilizado para variáveis categóricas e considera a fórmula de probabilidade de dois eventos independentes, e com a mesma probabilidade de ocorrência, acontecerem simultaneamente. Castro e Ferrari (2016, p. 264) esclarecem que o cálculo de Kappa é dado por:

$$K = \frac{Pr(a) - Pr(e)}{1 - Pr(e)} \tag{4}$$

$$Pr(a) = \frac{VP - VN}{N} \tag{5}$$

$$Pr(e) = \left(\frac{VP + FN}{N} \times \frac{VP + FN}{N}\right) + \left(\frac{FP + VN}{N} \times \frac{FN + VN}{N}\right) \tag{6}$$

Onde:

$P(a) =$ pares de observações concordantes, isto é, a acurácia;

$P(e) =$ probabilidade hipotética de concordância ao acaso;

$VP =$ verdadeiros positivos;

$VN =$ verdadeiros negativos;

$FN =$ falsos negativos;

$FP =$ falsos positivos;

$N =$ total de observações.

Os valores de Kappa, geralmente, variam entre 0 e 1, podendo haver valores negativos. Como regra geral, Kappa acima de 0,8 se traduz num bom nível de

concordância, quase perfeito; entre 0,67 e 0,8 relevante e; abaixo de 0,67 que a classificação realizada é duvidosa (MANNING; RAGHAVAN; SCHÜTZE, 2009, p. 166).

4.6 SÍNTESE DOS RECURSOS UTILIZADOS

No Quadro 6 é apresentada uma síntese dos recursos utilizados no desenvolvimento do método, contemplando as fases e as etapas deste Capítulo 4.

QUADRO 6 – RECURSOS UTILIZADOS NO DESENVOLVIMENTO DO MÉTODO, OBJETOS E FUNCIONALIDADES ENVOLVIDOS

Equipamento				
Microcomputador laptop Intel (R) Core (TM) i3 – 1005G1, CPU 1.20 GHz, RAM 8.00 Gb; Sistema Operacional 64b, Windows 10				
Fase	**Etapa**	**Software**	**Objeto**	**Pacote / função**
Compreensão dos dados	Coleta e análise dos dados	Access, Excel	Planilhas e tabelas	Microsoft Office 365
Preparação dos dados	Processamento das demandas em linguagem natural	Access	Tabelas	Microsoft Office 365
		R	*Corpora*	tm package / tolower(), removePuctuation(), removeNumbers(), acen(), stopwords(), stemDocument()
	Representação de documentos	R	DTMs	tm package / DocumentTermMatrix(), WeithTfIdf()
		R	DTMs	RWeka package / NGramTokenizer(), RemoveSparseTerms()
Modelagem	Processamento	Weka	*Datasets*	Algoritmos J48, IBk e Naïve Bayes Multinomial
	Treinamento e teste	Weka	*Datasets*	Validação cruzada
Avaliação do modelo	Avaliação de desempenho	Weka	*Datasets*	Coeficiente Kappa, taxa de acerto e tempo de processamento

FONTE: A autora (2021)

Posterior à síntese dos recursos utilizados na pesquisa, segue-se para apresentação dos resultados.

5 RESULTADOS

Neste capítulo são apresentados os resultados obtidos com a investigação do processo de gestão da informação do atendimento da Central 156 e a aplicação da metodologia CRISP-DM para responder a questão da pesquisa: como classificar as demandas da Central 156 de Curitiba utilizando o aprendizado de máquina?

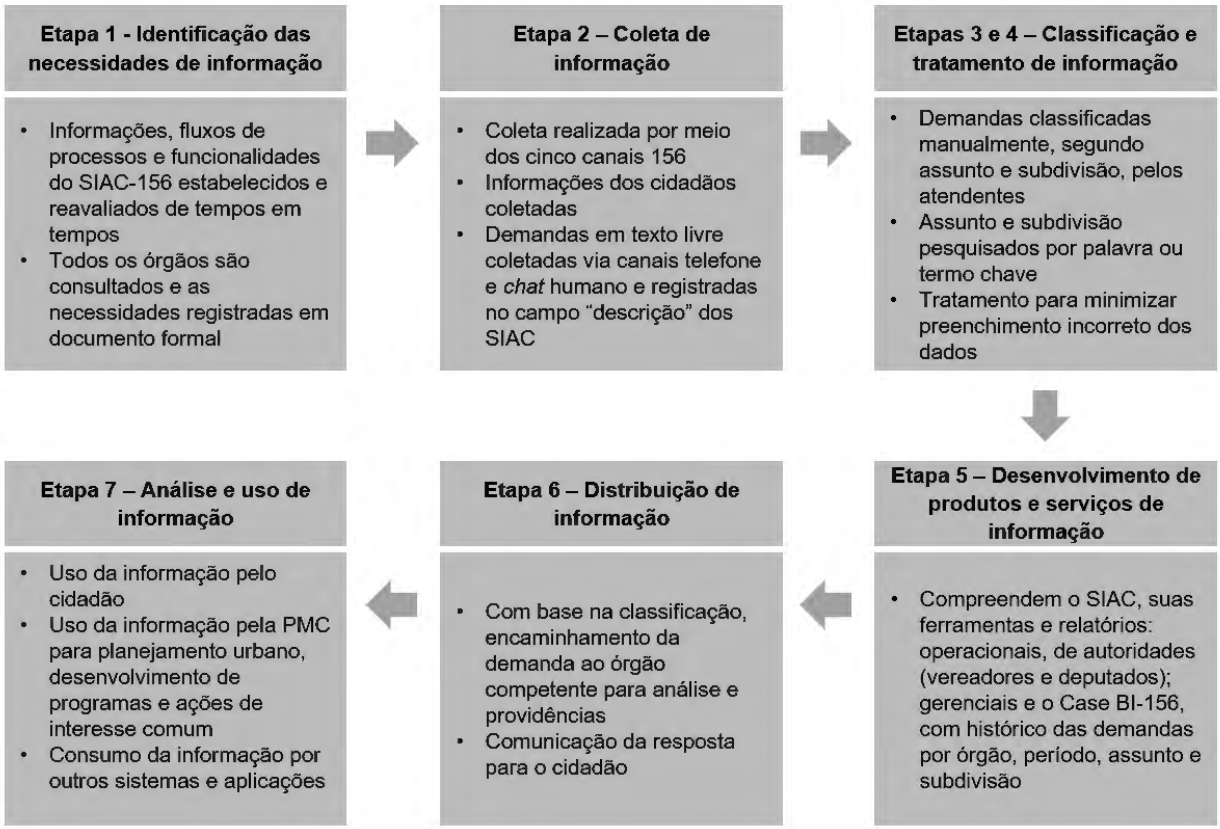
O capítulo está organizado seguindo a sequência de objetivos específicos da pesquisa. No tópico 5.1 demonstra-se o processo simplificado de gestão da informação do atendimento ao cidadão, realizado pela prefeitura via Central 156, correspondendo ao objetivo específico a). No tópico 5.2 são apresentados os resultados da aplicação do método, utilizando-se o processamento de linguagem natural e representando os documentos no modelo espaço vetorial, correspondendo às fases 2 e 3 do CRISP-DM, em atenção ao objetivo específico b) deste estudo. E no tópico 5.3 são apresentados os resultados referentes às fases 4 e 5 do CRISP-DM, modelagem e avaliação do modelo, correspondendo ao objetivo específico c). No tópico 5.4 é apresentado o modelo de classificação proposto.

5.1 PROCESSO DE GESTÃO DA INFORMAÇÃO DO ATENDIMENTO AO CIDADÃO DA CENTRAL 156

O fluxo do processo sintetizado de gestão da informação do atendimento ao cidadão, em consonância ao modelo de McGee e Prusak (1994, p. 107-126), compreendendo as etapas identificadas, é apresentado no diagrama da Figura 15.

A partir das tarefas que compõem as etapas do processo, é possível observar que a classificação e a distribuição das demandas aos órgãos responsáveis podem ser automatizadas por meio do processamento de linguagem natural e da aplicação dos algoritmos de classificação do aprendizado de máquina. Na classificação manual são utilizados termos e palavras-chave para busca e registro do assunto e da subdivisão de cada demanda. Da mesma forma, termos e palavras-chave, associados aos pesos de suas ocorrências no modelo espaço vetorial, possibilitam a classificação automática pelos algoritmos.

FIGURA 15 – FLUXO DO PROCESSO SIMPLIFICADO DE GESTÃO DA INFORMAÇÃO DO ATENDIMENTO AO CIDADÃO DA PMC, VIA CENTRAL 156



FONTE: A autora (2021)

O detalhamento do processo de gestão da informação do atendimento ao cidadão da PMC, realizado via Central 156, encontra-se no Apêndice 8.

5.2 RESULTADOS DO PROCESSAMENTO DE LINGUAGEM NATURAL E DA REPRESENTAÇÃO DOS DOCUMENTOS

Os experimentos foram realizados com unigramas e bigramas, de modo que se pudesse avaliar o desempenho dos algoritmos para as duas representações. A aplicação de dois limites para redução da esparsidade também adveio do interesse em avaliar o desempenho dos classificadores. Nesse caso, o propósito foi analisar a interferência do número de exemplos na classificação. Com o primeiro limite, procurou-se manter uma quantidade de palavras e de termos computacionalmente exequível, com o segundo, manter apenas a metade dessa quantidade. A Tabela 7 apresenta a dimensão das matrizes obtidas, antes e após a redução de esparsidade, com os valores limites utilizados.

TABELA 7 – DIMENSÃO DAS DTMs COM UNIGRAMAS E BIGRAMAS, ANTES E APÓS A REDUÇÃO DE ESPARSIDADE – *CORPUS* DA CLASSIFICAÇÃO POR ÓRGÃO

DTMs	Classificação por órgão	
	Limite	Dimensão (documentos x termos)
Unigrama inicial	NA	37.588 x 21.397
Unigrama – *dataset* maior	0,999	37.588 x 1.902
Unigrama – *dataset* menor	0,9975	37.588 x 1.004
Bigrama inicial	NA	37.588 x 258.802
Bigrama – *dataset* maior	0,999	37.588 x 2.118
Bigrama – *dataset* menor	0,9995	37.588 x 886

* NA – não se aplica.

FONTE: A autora (2021)

Observa-se, na Tabela 7, que o número de bigramas gerados ultrapassou 12 vezes o número de unigramas. Isso ocorreu devido à combinação das palavras que coocorrem nos documentos e que torna as matrizes ainda mais esparsas.

5.3 RESULTADOS DA APLICAÇÃO DOS ALGORITMOS DE APRENDIZADO DE MÁQUINA

Neste tópico são apresentados os resultados da classificação para unigramas e bigramas e posteriormente, a comparação entre os resultados obtidos com os experimentos.

5.3.1 Resultados da classificação com unigramas

O primeiro experimento consistiu na aplicação do algoritmo J48 para os dois *datasets* com unigramas. Na sequência, procedeu-se com a utilização dos algoritmos IBk e Naïve Bayes Multinomial, também para os dois *datasets*. Os resultados das métricas avaliadas constam na Tabela 8.

É possível observar que o coeficiente de Kappa e a taxa de acerto apresentaram valores muito semelhantes para os *datasets,* quando avaliados os algoritmos individualmente. O coeficiente de Kappa ficou acima de 0,8 para todos os algoritmos, indicando um bom nível de concordância entre a classificação predita e a obtida nos experimentos. As taxas de acerto atingiram cerca de 91% nos algoritmos J48 e Naïve Bayes Multinomial, para ambos *datasets*, e no IBk 84,6% para o *dataset*

menor. O tempo de processamento, no algoritmo J48, foi o dobro para o *dataset* maior, 2 horas e 13 minutos contra 58 minutos, no menor. Para o algoritmo Naïve Bayes Multinomial, os tempos foram praticamente iguais. Devido à falta de memória computacional, as métricas não foram calculadas no algoritmo IBk com o *dataset* maior.

TABELA 8 – RESULTADOS OBTIDOS NA CLASSIFICAÇÃO COM UNIGRAMAS

Algoritmo	**Métrica de desempenho**	**Unigrama – *dataset* maior**	**Unigrama – *dataset* menor**
J48	Coeficiente de Kappa	0,898	0,897
J48	Taxa de acerto	91,1%	91,0%
J48	Tempo de processamento	2h13min	58min
IBk	Coeficiente de Kappa	Não processou	0,823
IBk	Taxa de acerto		84,6%
IBk	Tempo de processamento		45min50s
Naïve Bayes Multinomial	Coeficiente de Kappa	0,900	0,894
Naïve Bayes Multinomial	Taxa de acerto	91,3%	90,8%
Naïve Bayes Multinomial	Tempo de processamento	6s	5s

FONTE: A autora (2021)

Nas Figuras 16 e 17 são apresentadas as matrizes de confusão para unigramas referentes aos dois *datasets*, considerando-se os três algoritmos avaliados. Como não foi possível executar o algoritmo IBk para o *dataset* maior, a matriz de confusão também não foi gerada.

Quanto ao *dataset* maior, analisando-se a Figura 16 e o Apêndice 9, para o algoritmo J48, é possível observar que a classe FAS apresentou a melhor taxa de acerto 98,4%, na qual apenas 67 exemplos, do total de 4.193, foram classificados incorretamente. A segunda melhor taxa foi da classe SMS, com 93,9%. Três classes ficaram com taxas de acerto abaixo de 90%, SMDT 83,6%, SGM 87,4% e SMMA 87,7%. SGM foi a classe mais atribuída aos exemplos pertencentes às outras classes.

FIGURA 16 – MATRIZES DE CONFUSÃO OBTIDAS NA CLASSIFICAÇÃO COM UNIGRAMAS – *DATASET* MAIOR

Algoritmo J48

	FAS	SGM	SMDT	SMMA	SMU	URBS	SMOP	SMS
FAS	4.126	15	9	16	3	7	0	17
SGM	12	4.017	152	145	93	64	50	65
SMDT	18	285	3.894	109	111	129	43	70
SMMA	12	250	113	4.523	96	52	19	90
SMU	7	104	115	54	4.200	16	10	47
URBS	5	83	101	36	11	4.302	10	43
SMOP	0	227	52	38	17	14	4.423	7
SMS	11	91	52	68	48	30	8	4.753

Algoritmo NaïveBayesMultinomial

	FAS	SGM	SMDT	SMMA	SMU	URBS	SMOP	SMS
FAS	4.060	15	21	28	7	12	0	50
SGM	1	4.131	102	71	154	37	28	74
SMDT	12	347	3.870	90	169	88	48	35
SMMA	10	205	52	4.733	86	9	23	37
SMU	1	89	97	82	4.243	9	13	19
URBS	11	43	124	23	30	4.300	19	41
SMOP	0	261	61	23	22	3	4.405	3
SMS	16	23	31	200	197	20	4	4.570

FONTE: A autora (2021)

Observando-se ainda a Figura 16 e o Apêndice 9, para o algoritmo Naïve Bayes Multinomial nota-se que, novamente, a classe FAS foi a que apresentou a melhor taxa de acerto 96,8%. A segunda melhor taxa foi da classe URBS, com 93,7%. Apenas duas classes obtiveram taxas abaixo de 90%, SMDT com 83,1% e SGM 89,8%. A SGM, no algoritmo Naïve Bayes Multinomial, também foi a classe mais atribuída aos exemplos de outras classes. As taxas para as demais classes podem ser consultadas no Apêndice 9.

A respeito do *dataset* menor, verifica-se, na Figura 17 e no Apêndice 10, para o algoritmo J48, que a classe FAS apresentou a melhor taxa de acerto 98,4%, seguida pela URBS, com 93,6%. Igualmente à classificação com unigramas, nesse algoritmo três classes ficaram com taxas de acerto inferiores a 90%, inclusive na mesma ordem, SMDT, com 83,0%, SGM, com 87,6% e SMMA com 88,8%. Com o algoritmo IBk, apenas duas classes obtiveram taxas de acerto superiores a 90%, FAS 96,4% e SMOP, com 91,1%. Três classes ficaram com taxas abaixo de 80%, SMDT 73,5%, URBS e SMS, com 78,8%.

Quanto ao algoritmo Naïve Bayes Multinomial, verifica-se, na Figura 17 e no Apêndice 10, que a classe FAS também foi a que apresentou a melhor taxa de acerto 96,6% seguida pela SMU, com 93,3%. Três foram as classes com taxas menores que 80%, SMDT 83,0%, SGM 88,6% e SMS 89,6%.

FIGURA 17 – MATRIZES DE CONFUSÃO OBTIDAS NA CLASSIFICAÇÃO COM UNIGRAMAS – *DATASET* MENOR

Algoritmo J48

	FAS	SGM	SMDT	SMMA	SMU	URBS	SMOP	SMS
FAS	4.124	14	12	12	4	6	0	21
SGM	8	4.030	158	132	91	64	51	64
SMDT	28	299	3.868	99	124	123	38	80
SMMA	7	221	96	4.576	96	52	22	85
SMU	5	103	137	39	4.205	11	10	43
URBS	7	86	122	15	11	4.296	7	47
SMOP	2	219	62	28	16	13	4.431	7
SMS	12	104	94	69	51	38	6	4.687

Algoritmo IBk

	FAS	SGM	SMDT	SMMA	SMU	URBS	SMOP	SMS
FAS	4.040	17	19	25	22	14	14	42
SGM	19	3.832	190	157	146	75	112	67
SMDT	32	371	3.426	203	157	179	136	155
SMMA	20	240	102	4.443	138	60	67	85
SMU	8	116	122	102	4.083	34	28	60
URBS	39	190	206	149	50	3.616	127	214
SMOP	6	224	78	47	23	34	4.354	12
SMS	44	255	133	237	107	230	68	3.987

Algoritmo NaïveBayesMultinomial

	FAS	SGM	SMDT	SMMA	SMU	URBS	SMOP	SMS
FAS	4.052	11	18	27	6	15	0	64
SGM	3	4.075	100	61	164	47	27	121
SMDT	11	347	3.869	84	146	107	44	51
SMMA	16	241	65	4.662	88	9	24	50
SMU	3	83	103	71	4.248	8	11	26
URBS	9	40	123	19	35	4.280	27	56
SMOP	0	260	73	24	26	2	4.391	2
SMS	20	26	42	224	184	26	4	4.535

FONTE: A autora (2021)

Assim como na classificação com unigramas, FAS foi a classe com taxas de acerto mais altas em todos os algoritmos utilizados e SGM a classe mais atribuída aos exemplos pertencentes às outras classes.

5.3.2 Resultados da classificação com bigramas

Na Tabela 9 constam os resultados referentes à taxa de acerto, ao coeficiente de Kappa e ao tempo de processamento, para os três algoritmos avaliados, considerando-se os bigramas, também para dois *datasets*.

TABELA 9 – RESULTADOS OBTIDOS NA CLASSIFICAÇÃO COM BIGRAMAS

Algoritmo	**Métrica de desempenho**	**Bigrama – *dataset* maior**	**Bigrama – *dataset* menor**
J48	Coeficiente de Kappa	0,834	0,827
J48	Taxa de acerto	85,5%	84,9%
J48	Tempo de processamento	2h38min	1h2min
IBk	Coeficiente de Kappa	Não processou	0,812
IBk	Taxa de acerto		83,6%
IBk	Tempo de processamento		25min20s
Naïve Bayes Multinomial	Coeficiente de Kappa	0,881	0,850
Naïve Bayes Multinomial	Taxa de acerto	89,6%	86,9%
Naïve Bayes Multinomial	Tempo de processamento	6s	3,5s

FONTE: A autora (2021)

Analisando-se a Tabela 9, percebe-se pouca variação entre os coeficientes de Kappa obtidos com os algoritmos, que ficaram entre 0,827 e 0,881, também acima de 0,8 como constatado na classificação com unigramas. Observa-se que praticamente não há diferença entre as taxas de acerto no algoritmo J48, as quais ficaram próximas de 85% para ambos os *datasets*. No algoritmo Naïve Bayes Multinomial ocorreu variação de 2,7% entre os *datasets*, sendo a taxa mais alta do *dataset* maior 89,6%. O tempo de processamento, no algoritmo J48, chegou quase a triplicar no *dataset* maior, 2 horas e 38 minutos contra 1 hora e 2 minutos, no menor. No Naïve Bayes Multinomial o tempo também foi praticamente o dobro, 6 segundos contra 3,5 segundos, no *dataset* menor.

As Figuras 18 e 19 apresentam as matrizes de confusão obtidas para os dois *datasets* com bigramas, considerando-se os três algoritmos estudados. Como ocorreu com o *dataset* maior de unigramas, não foi possível executar o algoritmo IBk por falta de memória computacional e, dessa maneira, a matriz de confusão não foi gerada.

FIGURA 18 – MATRIZES DE CONFUSÃO OBTIDAS NA CLASSIFICAÇÃO COM BIGRAMAS – *DATASET* MAIOR

Algoritmo J48

	FAS	SGM	SMDT	SMMA	SMU	URBS	SMOP	SMS
FAS	3.961	2	8	31	4	13	0	174
SGM	6	3.907	33	53	78	24	22	475
SMDT	6	222	3.374	65	45	83	13	901
SMMA	10	122	39	4.042	56	19	14	853
SMU	9	81	39	43	4.126	2	1	252
URBS	5	24	41	14	6	3.620	3	878
SMOP	4	197	29	22	5	15	4.241	265
SMS	7	17	30	52	21	24	2	4.908

Algoritmo NaiveBayesMultinomial

	FAS	SGM	SMDT	SMMA	SMU	URBS	SMOP	SMS
FAS	4.031	8	22	61	10	15	6	40
SGM	3	4.054	102	153	138	39	60	49
SMDT	29	250	3.879	205	104	80	53	59
SMMA	45	192	89	4.583	107	22	54	63
SMU	12	89	69	80	4.259	8	12	24
URBS	26	32	114	184	13	4.102	28	92
SMOP	7	230	86	64	16	8	4.352	15
SMS	16	39	52	373	113	36	15	4.417

FONTE: A autora (2021)

No que se refere ao *dataset* maior, analisando-se a Figura 18 e o Apêndice 11, para o algoritmo J48, observa-se que a taxa de acerto mais alta ficou com a classe SMS 97,0% e a segunda mais alta com a FAS 94,5%. Três classes ficaram com taxas de acerto abaixo de 80%, SMDT 71,3%, SMMA 78,4% e URBS 78,8%. SMS foi a classe mais atribuída aos exemplos pertencentes às outras classes.

Ainda com referência à Figura 18 e o Apêndice 11, para o algoritmo Naïve Bayes Multinomial, é possível observar que a classe FAS apresentou a melhor taxa de acerto 96,1%. A segunda melhor taxa foi da classe SMU, com 93,5% e as taxas mais baixas foram da SMDT 83,3% e da SMS 87,3%. Nesse algoritmo, SMMA foi a classe mais atribuída aos exemplos das outras classes.

Quanto ao *dataset* menor, para o algoritmo J48, é possível observar na Figura 19 e no Apêndice 12 que a classe SMS apresentou a melhor taxa de acerto 97,3%, seguida pela FAS, com 93,3%. Nesse algoritmo, três classes ficaram com taxas de acerto inferiores a 80%, SMDT com 70,5%, SMMA com 76,7% e URBS com 78,6% e SMS foi a classe mais atribuída aos exemplos pertencentes às outras classes. Com o algoritmo IBk, apenas uma classe ficou com taxa superior a 90%, a FAS 92,5%. Duas classes ficaram com taxas abaixo de 80%, SMDT 74,2% e URBS, com 79,4%. SMMA e SMS, com totais muito próximos, foram as classes mais atribuídas aos exemplos pertencentes às outras classes.

FIGURA 19 – MATRIZES DE CONFUSÃO OBTIDAS NA CLASSIFICAÇÃO COM BIGRAMAS – *DATASET* MENOR

Algoritmo J48

	FAS	SGM	SMDT	SMMA	SMU	URBS	SMOP	SMS
FAS	3.914	1	13	17	18	14	0	216
SGM	4	3.895	27	45	85	23	14	505
SMDT	9	220	3.285	68	50	98	9	920
SMMA	8	120	40	3.954	75	23	10	925
SMU	4	81	41	39	4.138	2	1	247
URBS	2	21	50	12	9	3.609	3	885
SMOP	5	201	32	21	8	14	4.200	297
SMS	6	15	32	33	22	24	7	4.322

Algoritmo IBk

	FAS	SGM	SMDT	SMMA	SMU	URBS	SMOP	SMS
FAS	3.877	19	41	108	9	31	2	106
SGM	25	3.781	134	268	101	64	59	216
SMDT	24	283	3.455	327	70	144	38	318
SMMA	35	186	114	4.299	61	54	29	377
SMU	15	104	106	157	4.018	20	12	121
URBS	23	63	181	272	12	3.644	8	388
SMOP	9	217	85	117	13	21	4.206	110
SMS	44	70	123	444	33	139	21	4.187

Algoritmo NaïveBayesMultinomial

	FAS	SGM	SMDT	SMMA	SMU	URBS	SMOP	SMS
FAS	3.971	6	19	116	7	27	7	40
SGM	7	3.904	87	305	138	41	59	57
SMDT	32	248	3.606	421	99	99	54	100
SMMA	88	164	69	4.557	100	25	63	89
SMU	25	94	77	117	4.171	12	27	30
URBS	22	25	86	397	6	3.887	37	131
SMOP	9	226	80	145	22	7	4.263	26
SMS	20	24	34	531	80	43	20	4.309

FONTE: A autora (2021)

Acerca do algoritmo Naïve Bayes Multinomial, FAS foi a classe que apresentou a melhor taxa de acerto 94,7% seguida pela SMU, com 91,6%. As taxas mais baixas foram observadas nas classes SMDT 77,4% e URBS 84,7%. Nesse algoritmo, SMMA foi a classe mais atribuída aos exemplos pertencentes às outras classes.

5.3.3 Comparação dos resultados

Com base nos resultados obtidos observa-se que houve pouca variação de desempenho quando comparadas as métricas entre os dois *datasets*, para o mesmo

algoritmo. Utilizando-se J48 para unigramas, o coeficiente de Kappa foi 0,898 no *dataset* maior e 0,897 no menor e para bigramas 0,834 no maior e 0,827 no menor. Com Naïve Bayes Multinomial, o coeficiente de Kappa para unigramas foi 0,900 no *dataset* maior e 0,894 no menor. A taxa de acerto no J48, com unigramas, foi 91,1% no *dataset* maior e 91,0% no menor, e com bigramas este algoritmo apresentou 85,5% no *dataset* menor e 84,9% no maior. Com Naïve Bayes Multinomial os valores foram 91,3% e 90,8% para unigramas e 89,6% e 86,9%, para os *datasets* maior e menor, respectivamente.

Quando comparados os tempos de processamento entre os *datasets*, a variação chegou a ser maior que o dobro no algoritmo J48, durando, na média, 1 hora e 23 minutos a mais para processar o *dataset* maior. No Naïve Bayes Multinomial praticamente não houve variação, visto que o algoritmo alcançou alto desempenho, sendo para unigramas 6 segundos e 5 segundos, e para bigramas 6 segundos e 3,5 segundos, respectivamente. No IBk não foi possível processar o *dataset* justamente devido à quantidade de atributos.

Quanto à representação dos documentos com n-gramas, verifica-se que, para o algoritmo IBk, os bigramas apresentaram desempenho discretamente inferior para o coeficiente de Kappa e para a taxa de acerto e, ainda, o tempo de processamento quase dobrou com bigramas. No algoritmo J48, essas três métricas também apresentaram desempenho inferior com bigramas. Para o *dataset* maior, o coeficiente de Kappa foi 0,898 com unigramas e 0,834 com bigramas, a taxa de acerto foi de 91,1% com unigramas e 85,5% para bigramas e o tempo de processamento foi 2 horas e 13 minutos e 2 horas e 38 minutos, respectivamente. E no Naïve Bayes Multinomial, o coeficiente de Kappa foi 0,900 com unigramas e 0,881 com bigramas, a taxa de acerto foi de 91,3% com unigramas e 89,6% para bigramas e o tempo de processamento foi o mesmo, 6 segundos.

No que se refere à taxa de acerto das classes, todas obtiveram percentuais acima de 70%. FAS foi a classe que apresentou desempenho superior em todos os experimentos, sendo a maior 98,4% obtida no algoritmo J48, com unigramas para o *dataset* maior. A exceção foi com o mesmo algoritmo e mesmo *dataset*, porém com bigramas, no qual a SMS apresentou resultado superior. Em contrapartida, SMDT foi a classe que apresentou o desempenho mais baixo em todos os experimentos, sendo a menor taxa no algoritmo J48 com bigramas para o *dataset* menor, ficando em 70,5%.

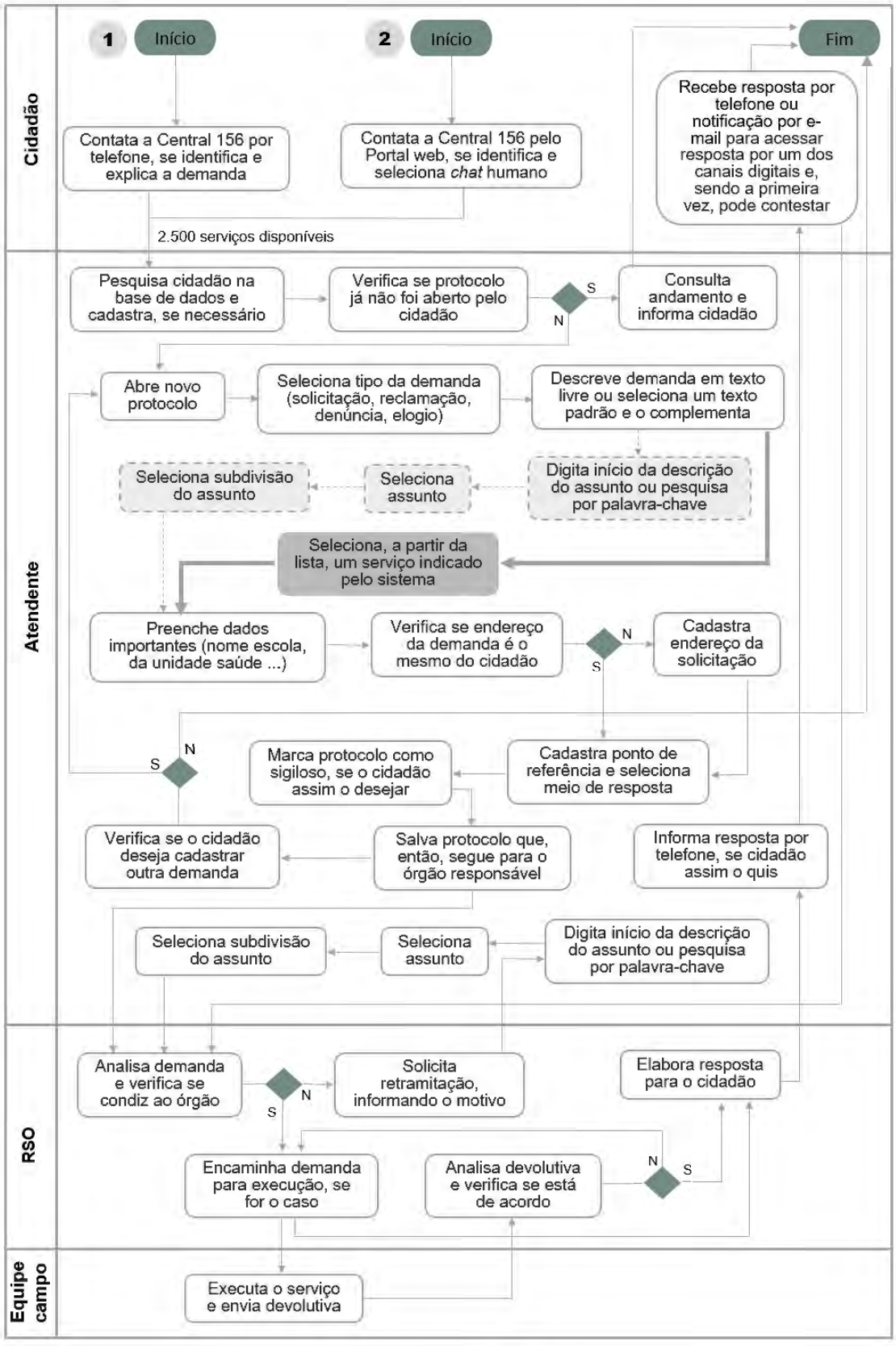
A classe FAS apresenta serviços caracterizados por palavras e termos exclusivos dessa classe, que podem ser observados na listagem dos serviços do Apêndice 6. Por outro lado, vários termos vinculados aos serviços da classe SMDT também estão associados aos serviços de outras classes, tais como “lâmpada” e “apagada” da classe SMOP; “estacionamento” da URBS e; “passeio” da SGM. Os serviços da classe SMDT também podem ser visualizados no Apêndice 6.

5.4 MODELO DE CLASSIFICAÇÃO PROPOSTO

De posse da descrição do processo de gestão da informação do atendimento ao cidadão, realizado via Central 156, e dos resultados para as métricas avaliadas, inclusive do coeficiente de Kappa, que indicou bons níveis de concordância entre a classificação realizada pelos atendentes e a obtida nos experimentos, foi possível redefinir o processo de classificação das demandas para as etapas realçadas na Figura 20.

Observa-se que o processo de classificação manual compreende as etapas de busca e seleção do assunto e da subdivisão, as quais podem ser automatizadas por meio da aplicação das técnicas abordadas neste estudo. De modo a proporcionar o processamento sem, contudo, esgotar a capacidade computacional, propõe-se um modelo de classificação que consiste na classificação em três estágios, iniciando pela classificação por órgão, depois por assunto e, então, por subdivisão. A Figura 21 apresenta as etapas do modelo proposto.

FIGURA 20 – FLUXO SIMPLIFICADO DE ATENDIMENTO AO CIDADÃO - CANAIS *CHAT* HUMANO E TELEFONE INCLUINDO A AUTOMATIZAÇÃO DA CLASSIFICAÇÃO DAS DEMANDAS



FONTE: A autora (2021)

FIGURA 21 – ETAPAS DO MODELO PROPOSTO DE CLASSIFICAÇÃO AUTOMÁTICA DAS DEMANDAS EM TRÊS ESTÁGIOS

FONTE: A autora (2021)

No modelo proposto a etapa de processamento, treinamento e teste é decomposta em três estágios, nas quais são gerados: um modelo de predição por órgão, n modelos de predição por assuntos, sendo um para cada órgão e n modelos de predição por subdivisões, sendo um para cada assunto existente. Na aplicação do modelo, a nova demanda passa pelas etapas de pré-processamento em linguagem natural e representação, sendo submetida à predição para identificação do órgão, depois para identificação do assunto desse órgão. Posteriormente, o sistema pode exibir ao atendente uma lista com os serviços que obtiveram maiores percentuais de classificação, de modo a auxiliá-lo no cadastramento do serviço referente à demanda. Outra possibilidade é o direcionamento automático da demanda para o órgão responsável, considerando-se o serviço com maior percentual de classificação.

6 CONSIDERAÇÕES FINAIS

Esta pesquisa teve por objetivo “propor um modelo de classificação automática para as demandas textuais da Central 156 de Atendimento ao Cidadão da Prefeitura de Curitiba, por meio de algoritmos de aprendizado de máquina”. Com vistas a atingir esse propósito, e pautando-se na literatura, foram definidos três objetivos específicos.

Quanto ao objetivo a) “descrever o processo de gestão da informação do atendimento ao cidadão, da Central 156”, procurou-se identificar, por meio da utilização do SIAC-156 e da leitura de seus manuais, as tarefas que compõem o processo de gestão da informação, no que concerne ao atendimento realizado pela prefeitura via Central 156. A partir desse objetivo, concluiu-se que a classificação e a distribuição das demandas, para os órgãos da PMC, são tarefas que podem ser automatizadas com utilização da inteligência artificial. Concluiu-se que palavras e termos chave são essenciais, sendo utilizados tanto pelos atendentes da Central 156 para busca e registro dos assuntos e das subdivisões das demandas, quanto para classificar automaticamente as demandas, ao constituírem o modelo espaço vetorial.

A condução da pesquisa experimental baseou-se no método CRISP-DM e a coleta dos dados foi realizada diretamente na base do 156, disponível no Portal de Dados Abertos da PMC, tendo como referência o ano de 2019. Como 98% das demandas estavam distribuídas entre oito órgãos, optou-se por manter essa configuração para estabelecer a população do estudo. Como havia grande desbalanceamento do número de registros entre os órgãos, empregou-se o método de reamostragem *undersampling*, considerado 95% de grau de confiança e 5% de margem de erro, segundo o órgão com menor número de registros. Posteriormente, devido à capacidade computacional, foi gerada uma amostra aleatória, considerando 99% de grau de confiança e 0,5% de margem de erro, totalizando 37.588 registros, conforme apresentado na Tabela 6.

Em atenção ao objetivo específico b) “submeter os textos das demandas ao processamento de linguagem natural, representando-os no modelo espaço vetorial”, foi realizado o pré-processamento dos textos, considerando-se o nível morfológico da linguagem, com aplicação de uma série de funções de tratamento de textos no idioma português. Permanecendo o conteúdo mais significativo, utilizando-se a abordagem *tokenization,* o conjunto de textos foi transformado na representação estruturada,

sendo a atribuição de pesos realizada por meio de ponderação TF-IDF. Com o propósito de diminuir a alta dimensionalidade e esparsidade da matriz, foram realizadas remoção de *stopwords,* conflação por *stemming* e eliminação de termos esparsos, sendo os valores limites para unigramas 0,999 e 0,9975, e para bigramas 0,999 e 0,9995, gerando quatro matrizes. O objetivo b) foi atingido com a conversão dessas matrizes em dois *datasets* para unigramas, de 1.902 e 1.004 palavras e dois para bigramas, com 2.118 e 886 termos, preparados para possibilitar a execução dos algoritmos.

Acerca do objetivo específico c) “aplicar algoritmos de aprendizado de máquina para classificação das demandas por órgão”, foram utilizados os algoritmos J48, IBk e Naïve Bayes Multinomial e avaliadas as métricas coeficiente de Kappa, taxa de acerto e tempo de processamento.

Concluiu-se que o tamanho dos *datasets* apresentou interferência ínfima no desempenho dos classificadores, com exceção do tempo de processamento para o algoritmo J48, que foi mais que o dobro para o *dataset* maior, tanto nos experimentos com unigramas quanto com bigramas. A respeito da representação, concluiu-se que unigramas proporcionaram melhor desempenho para o algoritmo J48 assim como para o Naïve Bayes Multinomial, não sendo possível avaliar o IBk uma vez que causou falta de memória computacional. Entre os classificadores, o melhor resultado foi obtido com o Naïve Bayes Multinomial. Concluiu-se, desse modo, que o melhor desempenho na classificação das demandas da Central 156 foi obtido com o algoritmo Naïve Bayes Multinomial, aplicado a unigramas no *dataset* maior, sendo as métricas atingidas 0,90 para o coeficiente de Kappa; 91,3% de taxa de acerto e 6 segundos de tempo de processamento.

A FAS foi a classe que apresentou o melhor desempenho em praticamente todos os experimentos, sendo a que constitui serviços caracterizados por palavras ou termos exclusivos dessa classe. A SMDT, em contrapartida, foi a classe que apresentou o pior desempenho, sendo também a classe que mais compartilha termos com outras classes. Assim, concluiu-se que a quantidade de termos comuns entre as classes tende a influenciar as taxas de acerto obtidas, isto é, quanto menos termos comuns, melhor a taxa de acerto, e quanto mais termos comuns, menor a taxa.

O modelo proposto para classificação das demandas (tópico 5.4), elaborado com base nos objetivos específicos que foram atingidos, decompõe a etapa de processamento, treinamento e teste em três estágios, nos quais são gerados: um

modelo de predição por órgão, n modelos de predição por assuntos, sendo um para cada órgão e n modelos de predição por subdivisões, sendo um para cada assunto existente. Na aplicação do modelo, a demanda a ser classificada passa inicialmente pelas etapas de pré-processamento em linguagem natural e representação, sendo submetida à predição, para identificação do órgão, depois do assunto relacionado a esse órgão e, então, para identificação da subdivisão desse assunto. Desse modo, são identificados os serviços mais prováveis para a nova demanda, de modo a auxiliar o atendente no cadastro. Outra possibilidade é direcionamento automático da demanda para o órgão responsável, considerando-se o serviço com maior percentual de classificação.

Assim sendo, considera-se que o objetivo geral desta pesquisa foi alcançado, uma vez que o modelo de classificação foi elaborado, conforme apresentado no tópico 5.4.

Acredita-se que, com os resultados promissores obtidos, a pesquisa possa contribuir com as discussões acerca do uso da inteligência artificial no âmbito municipal, no que se refere à classificação de demandas textuais, manifestas pelos cidadãos nas centrais 156 ou em portais *web*. A adoção das técnicas de processamento de linguagem natural e aprendizado de máquina possibilita o aprimoramento do processo de gestão da informação do atendimento ao cidadão, em razão de poder auxiliar os atendentes na classificação das demandas.

Ademais, o estudo contribui para o desenvolvimento de aplicações que classifiquem automaticamente as demandas registradas pelos cidadãos em texto livre, possibilitando maior participação dos cidadãos no direcionamento da oferta dos serviços públicos. Dessa maneira, o cidadão pode expressar e opinar sobre questões que não fazem parte do escopo de prioridades do governo, naquele momento. Por exemplo, as páginas *web* das consultas públicas, muitas vezes, contemplam apenas opções pré-definidas de uso de recursos, para seleção pelos cidadãos.

Identifica-se, também, potencial utilização do estudo em outros canais de atendimento ao cidadão. É o caso das demandas por informações públicas, abertas via Lei de Acesso à Informação. Na Prefeitura de Curitiba o cidadão redige a demanda e seleciona o órgão para envio. Por vezes, esse procedimento ocasiona tramitação do protocolo para o órgão indevido. Outro canal é o Fala Curitiba, que propicia ao cidadão, além da indicação de questões pré-definidas para inclusão na Lei Orçamentária Anual, o registro textual de demandas. Nesse contexto, o

processamento de linguagem natural e a classificação com aprendizado de máquina podem realizar a triagem das demandas, agilizando o atendimento. Desse modo, demandas de vários canais podem ser classificadas e agrupadas em relatórios, ampliando a qualidade da informação a ser utilizada no diagnóstico, planejamento e monitoramento da oferta dos serviços públicos.

6.1 LIMITAÇÕES DA PESQUISA

Os resultados da pesquisa quanto à compreensão do processo de gestão da informação do atendimento ao cidadão foram limitados à análise de manuais e do SIAC. Inicialmente, intencionava-se realizar uma observação participante, visando acompanhar as rotinas da Central 156, principalmente no tocante à classificação das demandas. Contudo, devido à pandemia causada pela Covid 19, essa atividade precisou ser cancelada.

Outra limitação foi quanto à capacidade de memória computacional, que impossibilitou a execução do algoritmo IBk quando utilizados bigramas para o *dataset* maior.

6.2 RECOMENDAÇÕES PARA TRABALHOS FUTUROS

Durante o desenvolvimento da pesquisa, constatou-se escassez de trabalhos contemplando a classificação automática de demandas textuais por serviços públicos, oportunizando a elaboração de estudos contemplando: a classificação automática de demandas textuais, considerando-se múltiplas classes; a utilização de outros algoritmos citados na literatura, como os baseados em redes neurais artificiais e nas máquinas de vetores de suporte e; a aplicação do modelo em projetos externos à prefeitura e que tenham como uma das suas ações as sugestões do público via texto livre, como o Projeto Curitiba 2035, disponível em *http://www.curitiba2035.org.br.*

REFERÊNCIAS

AGGARWAL, C. C.; ZHAI, C. A Survey of Text Classification Algorithms*. In:* AGGARWAL, C. C.; ZHAI, C. (ed.). **Mining Text Data.** New York: Springer, 2012. cap. 6, p.163-222. Disponível em: http://charuaggarwal.net/text-class.pdf. Acesso em: 17 maio 2019.

AGUNE, R. M.; CARLOS, J. A. Governo eletrônico e novos processos de trabalho. *In:* LEVY, E.; DRAGO, P. A. **Gestão Pública no Brasil Contemporâneo**, 2005. p. 1–16. Disponível em: https://governancaegestao.files.wordpress.com/2008/04/governo_eletronico_roberto_agune.pdf. Acesso em: 2 jun. 2020.

ALCANTARA, F. C. **Recuperação e classificação de informações provenientes da web e de redes sociais.** 2013. 119 f. Dissertação (Mestrado em Ciência, Gestão e Tecnologia da Informação) – Universidade Federal do Paraná, Curitiba, 2015. Disponível em: http://bdtd.ibict.br/vufind/Record/UFPR_e24772d23a53a3f4ed6d3cb2e906fa23. Acesso em: 18 set. 2020.

ANDRADE, M. B. D. E. **Proposta de inovação na gestão da informação na Companhia de Planejamento do Distrito Federal – Codeplan.** 2019. 55 f. Dissertação (Mestrado em Propriedade Intelectual e Transferência de Tecnologia para Inovação) – Universidade de Brasília, Brasília, 2019. Disponível em: https://repositorio.unb.br/handle/10482/38171. Acesso em: 18 set. 2020.

ANDROUTSOPOULOU, A. *et al.* Transforming the communication between citizens and government through AI-guided chatbots. **Government Information Quarterly,** v.36, n. 2, p. 358-367, set. 2018. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S0740624X17304008?via%3Dihub. Acesso em: 20 mar. 2020.

ARAMPATZIS, A. *et al.* Linguistically motivated information retrieval. **Encyclopedia of Library and Information Science,** v. 69, p. 1-24. 2000.

ARANHA, C. N. **Processamento automático para mineração de textos em português sob o enfoque da inteligência computacional.** 2007. 144 f. Tese (Doutorado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007. Disponível em: http://www.livrosgratis.com.br/ler-livro-online-73674/uma-abordagem-de-pre-processamento-automatico-para-mineracao-de-textos-em-portugues--sob-o-enfoque-da-inteligencia-computacional. Acesso em: 17 set. 2020.

ARANHA, C.; PASSOS, E. A Tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação,** v. 5, n. 2, p. 1-8, 2006. Disponível em: http://www.periodicosibepes.org.br/index.php/reinfo/article/view/171/66. Acesso em: 20 jun. 2019.

ATHENA TECH LLC. AI, Machine learning (ML) and Natural Language Processing (NLP). **My Blog.** 20 out. 2019. Disponível em: https://athenatech.tech/f/ai-machine-learning-ml-and-natural-language-processing-nlp. Acesso em: 10 ago. 2020.

AWAD, M.; KHANNA, R. **Efficient Learning Machines**: Theories, Concepts, and Applications for Engineers and System Designers. Apress, 2015. Disponível em: https://library.oapen.org/bitstream/handle/20.500.12657/28170/1001824.pdf?sequence=1. Acesso em: 12 nov. 2019.

BAGHDADI, Y*. et al.* Automatic classification of free-text medical causes from death certificates for reactive mortality surveillance in France. **International Journal of Medical Informatics,** v. 131, nov. 2019. Disponível em: https://doi.org/10.1016/j.ijmedinf.2019.06.022. Acesso em: 15 mar. 2020.

BARBA, P. Machine learning (ML) for Natural Language Processing (NLP). **Machine Learning.** 29 set. 2020. Disponível em: https://www.lexalytics.com/lexablog/machine-learning-natural-language-processing. Acesso em: 15 jan. 2021.

BARROS, A. J. DA S.; LEHFELD, N. A. DE S. **Fundamentos de metodologia científica.** 3. ed. São Paulo: Pearson Prentice Hall, 2007.

BERTI, C. B. **Modelo preditivo de situações como apoio à consciência situacional e ao processo decisório em sistemas de resposta à emergência.** 2017. 153 f. Tese (Doutorado em Ciência da Computação) – Universidade Federal de São Carlos, São Carlos, 2017. Disponível em: https://repositorio.ufscar.br/handle/ufscar/10119?show=full. Acesso em: 21 set. 2020.

BRAMER, M. **Principles of data mining.** 3rd ed. London: Springer, 2016.

BRASIL. Senado Federal. **Projeto de Lei nº 5.691, de 2019.** Institui a Política Nacional de Inteligência Artificial. Brasília, 2019. Disponível em: https://www25.senado.leg.br/web/atividade/materias/-/materia/139586. Acesso em: 14 abr. 2021.

BRASIL. **Lei nº 13.460, de 26 de junho de 2017.** Dispõe sobre participação, proteção e defesa dos direitos do usuário dos serviços públicos da administração pública. Brasília, DF: Presidência da República, [2017]. Disponível em: http://www.planalto.gov.br/ccivil_03/_ato2015-2018/2017/lei/l13460.htm. Acesso em: 2 jun. 2020.

BRASIL. [Constituição (1988)]. **Constituição da República Federativa do Brasil:** texto constitucional promulgado em 5 de outubro de 1988, com as alterações adotadas pelas emendas constitucionais nos 1/1992 a 99/2017, pelo Decreto legislativo nº 186/2008 e pelas emendas constitucionais de revisão nos 1 a 6/1994. 53. ed. Brasília: Câmara dos Deputados, Edições Câmara, 2018.

CASTRO, L. N.; FERRARI, D. G. **Introdução à mineração de dados:** conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

CHOO, C. W. **A organização do conhecimento:** como as organizações usam a informação para criar significado, construir conhecimento e tomar decisões. Tradução Eliana Rocha. São Paulo: Editora Senac São Paulo, 2003.

CHOO, C. W. Information Management. **Faq's.** Disponível em: http://choo.fis.utoronto.ca/Imfaq/. Acesso em: 17 jan. 2021.

CHOWDHURY, G. G. Natural language Processing. **Annual Review of Information Science and Technology,** v. 37, p. 51-89, 2003.

CONDURÚ, M. T.; PEREIRA, J. A. R. Gestão da informação em saneamento básico no Estado do Pará sob o enfoque do ciclo informacional. **Engenharia Sanitária e Ambiental,** v. 22, n. 6, p. 1225–1232, nov./dez. 2017. Disponível em: http://repositorio.ufpa.br/jspui/bitstream/2011/11372/1/Artigo_GestaoInformacaoSaneamento.pdf. Acesso em: 15 fev. 2021.

CURITIBA. **Lei nº 15.798, de 22 de dezembro de 2020.** Estima a Receita e fixa a Despesa do Município de Curitiba para o exercício financeiro de 2021. Curitiba: Câmara Municipal, [2020]. Disponível em: https://legisladocexterno.curitiba.pr.gov.br/AtosConsultaExterna.aspx. Acesso em: 5 maio. 2021.

DAVENPORT, T. H.; PRUSAK, L. **Ecologia da informação:** porque só a tecnologia não basta para o sucesso na era da informação. Tradução Bernadette Siqueira Abrão. São Paulo: Futura, 1998.

DAVENPORT, T. H.; PRUSAK, L. Working Knowledge: How Organizations Manage What They Know. **Ubiquity,** v. 2000, ago. 2000. Disponível em: https://ubiquity.acm.org/article.cfm?id=348775. Acesso em: 15 set. 2019.

DETLOR, B. Information Management. **International Journal of Information Management,** v. 30, n. 2, p. 103–108, 2010.

DIIRR, B.; ARAUJO, R. M. DE; CAPPELLI, C. Conversas sobre serviços públicos. *In:* SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO, 7., 2011, Salvador. **Anais** [...]. Salvador: Universidade Federal da Bahia, 2011, p. 396-407. Disponível em: https://sol.sbc.org.br/index.php/sbsi/article/view/14593/14440. Acesso em: 15 jun. 2020.

EXAMTIME. **GoConqr.** 2021. Disponível em: https://www.goconqr.com/pt-BR. Acesso em: 25 jun. 2021.

EYHERAMENDY, S.; LEWIS, D. D.; MADIGAN, D. On the Naive Bayes model for text categorization. *In:* INTERNATIONAL WORKSHOP ON ARTIFICIAL INTELLIGENCE AND STATISTICS, 9., Key West, 2003. **Proceedings** [...]. PMLR, 2003, p. 93-100. Disponível em: https://proceedings.mlr.press/r4/eyheramendy03a/eyheramendy03a.pdf. Acesso em: 12 jul. 2020.

FANG, Z. E-government in digital era: concept practice and development. **International Journal of The Computer, The Internet and Management,** v. 10, n. 2, p. 1-22, 2002. Disponível em: http://www.ijcim.th.org/past_editions/2002V10N2/article1.pdf. Acesso em: 5 jul. 2020.

FEINERER, I.; HORNIK, K. **Package "tm".** 2019. Disponível em: https://cran.r-project.org. Acesso em: 30 dez. 2019.

FEINERER, I.; HORNIK, K.; MEYER, D. Text mining infrastructure in R. **Journal of Statistical Software,** v. 25, n. 5, p. 1-54, 2008. Disponível em: https://www.jstatsoft.org/article/view/v025i05. Acesso em: 10 set. 2020.

FELDMAN, R.; SANGER, J. **The Text Mining Handbook:** Advanced Approaches in Analyzing Unstructured Data. Cambridge University Press, 2007.

GALLO, A. J. DE M. Gestão estratégica da informação no ambiente do governo digital. **Revista Brasileira de Biblioteconomia e Documentação,** v. 6, n. 2, p. 3-19, 2010. Disponível em: https://rbbd.febab.org.br/rbbd/article/view/126. Acesso em: 15 jan. 2021.

GIL, A. C. **Métodos e técnicas de pesquisa social.** 6. ed. São Paulo: Atlas, 2008.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining:** um guia prático. Rio de Janeiro: Elsevier, 2005.

GONZALEZ, M.; LIMA, V. L. S. DE. Recuperação de informação e processamento da linguagem natural. *In:* CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO, 23., 2003, Porto Alegre. **Anais** [...]. Porto Alegre: Pontifícia Universidade Católica do Rio Grande do Sul, 2003, p. 347-395. Disponível em: https://www.marilia.unesp.br/Home/Instituicao/Docentes/EdbertoFerneda/mri-06---gonzales-e-lima-2003.pdf. Acesso em: 10 ago. 2020.

HAN, E.; KARYPIS, G.; KUMAR, V. **Text categorization using weight adjusted k-nearest neighbor classification.** Minneapolis: University of Minnesota, 1999.

HERINGER, L. P. *et al.* Governo eletrônico e gestão do relacionamento com o cidadão (CIRM): estudo de caso da *fan page* da Secretaria de Estado de Saúde de Minas Gerais. ENCONTRO DE ADMINISTRAÇÃO DE INFORMAÇÃO (ENADI), 6, 2017. **Anais** [...]. Curitiba: Grupo de Pesquisa Estado, Democracia e Administração Pública, 2017, p. 1-8. Disponível em: https://www.academia.edu/37337639/Governo_Eletrônico_e_Gestão_do_Relacionamento_com_o_Cidadão_CIRM_Estudo_de_Caso_da_Fan_Page_da_Secretaria_de_Estado_de_Saúde_de_Minas_Gerais. Acesso em: 5 jul. 2020.

IBM. **Guia do IBM SPSS Modeler CRISP-DM**. 2015. Disponível em: ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/17.1/br_po/ModelerCRISPDM.pdf. Acesso em: 11 maio 2019.

ICI. INSTITUTO DAS CIDADES INTELIGENTES. **Sistema integrado de atendimento ao cidadão:** Manual de instrução do cadastrador. Curitiba: ICI, 2019a.

ICI. INSTITUTO DAS CIDADES INTELIGENTES. **Sistema integrado de atendimento ao cidadão:** Manual de instrução do responsável pelo serviço no órgão. Curitiba: ICI, 2019b.

ICI. INSTITUTO DAS CIDADES INTELIGENTES. **Central 156:** Estatísticas & indicadores, dez. 2019. Curitiba: ICI, 2019c. Disponível em: http://multimidia.central156.org.br/2020/8/pdf/00000335.pdf. Acesso em: 20 set. 2020.

ICI. INSTITUTO DAS CIDADES INTELIGENTES. **Relatório tabela de serviço.** Curitiba: ICI, 2021a.

ICI. INSTITUTO DAS CIDADES INTELIGENTES. **Sistema integrado de atendimento ao cidadão (SIAC-156).** Curitiba: ICI, 2021b.

ICI. INSTITUTO DAS CIDADES INTELIGENTES. **Organização social.** Curitiba: ICI, 2021c. Disponível em: https://www.ici.curitiba.org.br/conteudo/organizacao-social/5. Acesso em: 22 set. 2020.

ICI. INSTITUTO DAS CIDADES INTELIGENTES. **Central 156 de Curitiba.** Curitiba: ICI, 2021d. Disponível em: http://www.central156.org.br/. Acesso em: 22 set. 2020.

ICI. INSTITUTO DAS CIDADES INTELIGENTES. **Aplicativo Curitiba 156.** Curitiba: ICI, 2021e.

IPPUC. INSTITUTO DE PESQUISA E PLANEJAMENTO URBANO DE CURITIBA. **Histórico do 156.** Curitiba: IPPUC, 1991.

IPPUC. INSTITUTO DE PESQUISA E PLANEJAMENTO URBANO DE CURITIBA. **SEUC - Quantitativo de equipamentos urbanos municipais**. Curitiba: IPPUC, 2021.

JACKSON, P.; MOULINIER, I. **Natural Language Processing for Online Applications:** Text Retrieval, Extraction and Categorization. John Benjamins Publishing, 2002.

JUSTEN FILHO, M**. Curso de direito administrativo.** 13. ed. São Paulo: Editora Revista dos Tribunais, 2018.

KANO, E.; FUJITA, Y.; TSUDA, K. A method of extracting and classifying local community problems from citizen-report data using text mining. **Procedia Computer Science,** v. 159, p. 1347-1356, 2019. Disponível em: https://doi.org/10.1016/j.procs.2019.09.305. Acesso em: 7 jun. 2020.

KUZIEMSKI, M.; MISURACA, G. AI governance in the public sector: three tales from the frontiers of automated decision-making in democratic settings. **Telecommunications Policy,** v. 44, n. 6, p. 1-13, 2020. Disponível em: https://doi.org/10.1016/j.telpol.2020.101976. Acesso em: 4 jun. 2019.

LAROSE, D. T.; LAROSE, C. D. **Discovering knowledge in data:** an introduction to data mining. 2nd ed. New Jersey: Wiley, 2014.

LIDDY, E. D. Natural language processing. *In:* DRAKE M.; MAACK, M. N. (ed.). **Encyclopedia of Library and Information Science**, 2nd Ed. New York: Marcel Decker, Inc, 2001.

LIMA, G. Â. B. DE O. Modelos de categorização: apresentando o modelo clássico e o modelo de protótipos. **Perspectivas em Ciência da Informação,** v. 15, n. 2, p. 108-122, 2010. Disponível em: https://www.scielo.br/j/pci/a/Mzmh4hhnBMt5ym3zjwcWCLG/?format=pdf&lang=pt. Acesso em: 12 set. 2020.

LIU, A. Y. **The effect of oversampling and undersampling on classifying imbalanced text datasets.** 2004. 57 f. Dissertação (Mestrado Ciências em Engenharia) – University of Texas, Austin, 2004.

LOH, S. **BI na era do big data para cientistas de dados:** indo além de cubos e dashboards na busca pelos porquês, explicações e padrões. 1. ed. Porto Alegre, 2014.

LOPES, I. L.; SANTOS, F. A. O.; PINHEIRO, C. A. M. **Inteligência Artificial.** 1. ed. Rio de Janeiro: Elsevier, 2014.

MA, L.; ZHENG, Y. National e-government performance and citizen satisfaction: a multilevel analysis across European countries. **International Review of Administrative Sciences,** v. 85, n. 3, p. 506-526, 2019.

MANNING, C. D.; RAGHAVAN, P.; SCHÜTZE, H. **An introduction to Information Retrieval.** Cambridge: Cambridge University Press, 2009.

MANNING, C. D.; SCHÜTZE, H. **Foundations of Statistical Natural Language Processing**. Cambridge: MIT Press, 1999.

MARCHIORI, P. Z. A ciência e a gestão da informação: compatibilidades no espaço profissional. **Ciência da Informação,** v. 31, n. 2, p. 72-79, 2002. Disponível em: http://revista.ibict.br/ciinf/article/view/962/999. Acesso em: 4 maio 2021.

MARTINS, C. A.; MONARD, M. C.; MATSUBARA, E. T. Reducing the Dimensionality of Bag-of-Words Text Representation Used by Learning Algorithms. *In:* IASTED INTERNATIONAL CONFERENCE ON ARTIFICIAL INTELLIGENCE AND APPLICATIONS, 3., 2003. **Proceedings** [...]. 2003, p. 228-233. Disponível em: https://www.researchgate.net/profile/Edson-Matsubara/publication/228983901_Reducing_the_dimensionality_of_bag-of-words_text_representation_used_by_learning_algorithms/links/5501a1410cf231de076a9502/Reducing-the-dimensionality-of-bag-of-words-text-represen. Acesso em: 8 mar. 2021.

MCCALLUM, A.; NIGAM, K. A comparison of event models for Naive Bayes text classification. *In:* AAAI-98 WORKSHOP ON LEARNING FOR TEXT

CATEGORIZATION, 1998, Madison. **Proceedings** [...]. Madison: AAAI Press,1998, p. 41-48. Disponível em: https://www.cs.cmu.edu/~knigam/papers/multinomial-aaaiws98.pdf. Acesso em: 14 abr. 2021.

MCCARTHY, J.; FEIGENBAUM, E. A. Arthur Samuel (in memoriam): pioneer in machine learning. **AI Magazine,** v. 11, n. 3, p. 10-11, 1990. Disponível em: https://ojs.aaai.org//index.php/aimagazine/article/view/840. Acesso em: 15 ago. 2020.

MCGEE, J.; PRUSAK, L. **Gerenciamento estratégico da informação:** aumente a competitividade e a eficiência de sua empresa utilizando a informação como uma ferramenta estratégica. Tradução: Astrid Beatriz de Figueiredo. Rio de Janeiro: Campus, 1994.

MEHR, H.; ASH, H.; FELLOW, D. Artificial intelligence for citizen services and government. **Ash Cent. Democr. Gov. Innov. Harvard Kennedy Sch.,** August, p. 1-12, 2017.

MEIJER A. J.; CURTIN D.; HILLEBRANDT M. Open government: connecting vision and voice. **International Review of Administrative Sciences,** v. 78, n. 1, p. 10–29, 2012. Disponível em: https://journals.sagepub.com/doi/pdf/10.1177/0020852311429533. Acesso em: 27 jun. 2019.

MENDES JUNIOR, R. Crowdsourcing e machine learning: uma revisão sistemática com discussão do uso para a participação pública dos cidadãos. **Inteligência Artificial:** 3º Grupo de Pesquisa do ITS, Rio de Janeiro**,** 2018. Disponível em: https://itsrio.org/wp-content/uploads/2019/03/Ricardo-Mendes.pdf. Acesso em: 13 jul. 2020.

MEZZAROBA M. P.; BIER C. A. Revisão sistemática da literatura sobre democracia eletrônica e governo eletrônico. **Conpedi Law Review**, v. 1, n. 9, p. 208-233, 2016. Disponível em: https://www.indexlaw.org/index.php/conpedireview/article/view/3379/2896. Acesso em: 14 set. 2020.

MINISTÉRIO DAS COMUNICAÇÕES. Serviços de utilidade pública e de emergência (SUP). **Agência Nacional de Telecomunicações.** 2015. Disponível em: https://www.gov.br/anatel/pt-br/regulado/numeracao/codigos-nacionais/servicos-de-utilidade-publica-e-de-emergencia. Acesso em: 20 mar. 2020.

MITCHELL, T. M. Machine Learning. New York: McGraw-Hill Education, 1997.

MONARD, M. C.; BARANAUSKAS, J. A. Conceitos sobre aprendizado de máquina. In: REZENDE S. O. (CCC) **Sistemas Inteligentes:** Fundamentos e Aplicações. Barueri: Editora Manole, 2003. cap 4, p. 39-56. Disponível em: https://dcm.ffclrp.usp.br/~augusto/publications/2003-sistemas-inteligentes-cap4.pdf. Acesso em: 3 mar. 2021.

MONTEIRO, L. L. **Mensagens textuais no canal de atendimento do portal IBGEANDO:** obtendo insumos para a tomada de decisão utilizando mineração de textos. 2017. 163 f. Dissertação (Mestrado em Ciência da Informação) – Universidade Federal Fluminense, Niterói, 2017. Disponível em: https://app.uff.br/riuff/handle/1/10875. Acesso em: 29 set. 2020.

NAM, T. New Ends, New Means, but Old Attitudes: Citizens' Views on Open Government and Government 2.0. *In:* HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 44., 2011, Kauai. **Proceedings** [...]. New York: IEEE Computer Society, 2011, p. 1-10.

NOGUEIRA, E. D. A. **Análise de *brand equity* sob a perspectiva do consumidor nas mídias sociais por meio da mineração de opinião e análise de redes sociais.** 2015. 235 f. Dissertação (Mestrado em Ciência, Gestão e Tecnologia da Informação) – Universidade Federal do Paraná, Curitiba, 2015. Disponível em: https://acervodigital.ufpr.br/handle/1884/41257. Acesso em: 22 set. 2020.

PINHO, J. A. G. *et al.* Democracia digital na área de administração: um levantamento da construção do campo no Brasil. **Cadernos Gestão Pública e Cidadania,** v. 24, n. 78, p. 1-31, 2019. Disponível em: http://bibliotecadigital.fgv.br/ojs/index.php/cgpc/article/view/73630/76299. Acesso em: 21 abr. 2021.

PMC. PREFEITURA MUNICIPAL DE CURITIBA. 156 - Histórico - Base de Dados. **Portal de Dados Abertos.** 2019a. Disponível em: http://dadosabertos.c3sl.ufpr.br/curitiba/156/. Acesso em: 15 mar. 2020.

PMC. PREFEITURA MUNICIPAL DE CURITIBA. 156 - Dicionário - Base de Dados. **Portal de Dados Abertos.** 2019b. Disponível em: https://www.curitiba.pr.gov.br/dadosabertos/. Acesso em: 15 mar. 2020.

PMC. PREFEITURA MUNICIPAL DE CURITIBA. Relação de Servidores / Empregados Ativos. **Portal da transparência.** 2020. Disponível em: http://multimidia.transparencia.curitiba.pr.gov.br/funcionarios/12_RELACAO_DE_SERVIDORES_ATIVOS_DEZEMBRO2020.pdf. Acesso em: 5 fev. 2021.

PMC. PREFEITURA MUNICIPAL DE CURITIBA. Dados da Cidade de Curitiba. **Perfil da Cidade de Curitiba.** 2021a. Disponível em: https://www.curitiba.pr.gov.br/conteudo/perfil-da-cidade-de-curitiba/174. Acesso em: 5 fev. 2021.

PMC. PREFEITURA MUNICIPAL DE CURITIBA. Secretarias. **Portal da Prefeitura de Curitiba.** 2021b. Disponível em: https://www.curitiba.pr.gov.br/. Acesso em: 5 fev. 2021.

POLLETTINI, J. T. **Avaliação de mecanismos de suporte à tomada de decisão e sua aplicabilidade no auxílio à priorização de casos em regulações de urgências e emergências.** 2016. 97 f. Tese (Doutorado em Ciências Médicas) – Faculdade de Medicina de Ribeirão Preto, Universidade de São Paulo, Ribeirão

Preto, 2016. Disponível em: https://teses.usp.br/teses/disponiveis/17/17138/tde-30032017-101723/pt-br.php. Acesso em: 18 set. 2020.

PONJUÁN DANTE, G. **Gestión de la Información:** dimensiones e implementación para el éxito organizacional. Rosario: Nuevo Paradigma, 2004.

PRADO, E. P. V. *et al.* Iniciativas de governo eletrônico: análise das relações entre nível de governo e características dos projetos em casos de sucesso. **Revista Eletrônica de Sistemas de Informação,** v. 10, n. 1, p. 1-22, 2011. Disponível em: http://www.periodicosibepes.org.br/index.php/reinfo/article/view/793/pdf. Acesso em: 14 jul. 2020.

PRATI, R. C. **Novas abordagens em aprendizado de máquina para a geração de regras, classes desbalanceadas e ordenação de casos.** 2006. 191 f. Tese (Doutorado em Ciências de Computação e Matemática Computacional) – Universidade Federal de São Carlos, São Carlos, 2006. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-01092006-155445/pt-br.php. Acesso em: 14 set. 2020.

RICHARDSON, R. J. **Pesquisa social:** métodos e técnicas. 3. ed. São Paulo: Atlas, 2012.

RINGOLD, D. *et al.* **Citizens and Service Delivery:** Assessing the Use of Social Accountability Approaches in the Human Development Sectors. Washington DC: World Bank Publications, 2012.

RODRIGUES, T. R. G.; PARRÃO, J. A. A necessidade da implantação da gestão da informação no Adra/Cadeca. **Revista Seminário Integrado,** v. 11, n. 11, p. 1–17, 2017.

ROSSI, R. G. **Classificação automática de textos por meio de aprendizado de máquina baseado em redes.** 2015. 315 f. Tese (Doutorado em Ciências de Computação e Matemática Computacional) – Universidade Federal de São Carlos, São Carlos, 2015. Disponível em: https://www.teses.usp.br/teses/disponiveis/55/55134/tde-05042016-105648/pt-br.php. Acesso em: 16 out. 2020.

RUSSELL, S.; NORVIG, P. **Inteligência Artificial.** Tradução: Regina Célia Simille de Macedo. 3. ed. Rio de Janeiro: Elsevier, 2013.

SANTOS, K. B. C. **Categorização de textos por aprendizagem de máquina.** 2020. 85 f. Dissertação (Mestrado em Modelagem Computacional de Conhecimento) - Universidade Federal de Alagoas, Maceió, 2019. Disponível em: https://keilabcs.github.io/publications/dissertacao.pdf. Acesso em: 26 set. 2020.

SANTOS, P. M.; ROVER, A. J. Direito à informação e à participação: uma avaliação das ferramentas dispostas nos portais de governo estaduais. **Informação & Sociedade,** v. 28, n. 1, p. 219-244, 2018. Disponível em: https://periodicos.ufpb.br/index.php/ies/article/view/38171/19706. Acesso em: 10 jun. 2020.

SCHELLONG, A. CRM in the public sector: towards a conceptual research framework. *In:* NATIONAL CONFERENCE ON DIGITAL GOVERNMENT RESEARCH, 2005, Atlanta. **Proceedings** [...]. Wadern: Schloss Dagstuhl, 2005. p.326–332.

SCHMIDTHUBER, L*. et al.* The emergence of local open government: determinants of citizen participation in online service reporting. **Government Information Quarterly,** v. 34, n. 3, p. 457-469, 2017. Disponível em: https://doi.org/10.1016/j.giq.2017.07.001. Acesso em: 9 set 2019.

SCHNEIDER, K.-M. Techniques for improving the performance of Naive Bayes for text classification*. In:* Gelbukh A. (ed.). **Computational Linguistics and Intelligent Text Processing.** Heidelberg: Springer, 2005. p. 682-693.

SEBASTIANI, F. Machine learning in automated text categorization. **ACM Computing Surveys,** v. 34, n. 1, p. 1-47, 2002. Disponível em: https://arxiv.org/pdf/cs/0110053.pdf. Acesso em: 11 out. 2020.

SETZER, V. W. Dado, informação, conhecimento e competência. **DataGramaZero Revista de Ciência da Informação,** v. 28, n. 0, p. 1-14, 1999. Disponível em: https://brapci.inf.br/index.php/res/download/45629. Acesso em: 5 mar. 2021.

SGM. SECRETARIA DO GOVERNO MUNICIPAL. **Histórico do sistema de atendimento ao cidadão.** Curitiba: SGM, 2015.

SILVA, A. B. M. DA; RIBEIRO, F. A gestão da informação na administração pública. **Interface Administração Pública,** v. 161, n. 50, p. 32-39, 2009.

SILVA, M.; SOUZA, R. Fundamentos em processamento de linguagem natural: uma proposta para extração de bigramas. **Encontros Bibli:** Revista Eletrônica de Biblioteconomia e Ciência da Informação, v. 19, n. 40, p. 1-32, 2014. Disponível em: https://www.researchgate.net/publication/274915215_Fundamentos_em_processamento_de_linguagem_natural_uma_proposta_para_extracao_de_bigramas. Acesso em: 10 ago. 2020.

SORDI, J. O. DE. **Administração da informação:** fundamentos e práticas para uma nova gestão do conhecimento. São Paulo: Saraiva, 2008.

SUN, P.; KU, C.; SHIH, D. An implementation framework for e-government 2.0. **Telematics and Informatics,** v. 32, n. 3, p. 504-520, 2015. Disponível em: http://dx.doi.org/10.1016/j.tele.2014.12.003. Acesso em: 19 jun. 2020.

TAN, C.; WANG, Y.; LEE, C. The use of bigrams to enhance text categorization. **Information Processing & Management**, v. 38, n. 4, p. 529-546, 2002.

TAVANA, M.; ZANDI, F.; KATEHAKIS, M. N. A hybrid fuzzy group ANP-TOPSIS framework for assessment of e-government readiness from a CiRM perspective. **Information & Management,** v. 50, n. 7, p. 383-397, 2013. Disponível em: http://dx.doi.org/10.1016/j.im.2013.05.008. Acesso em: 9 fev. 2021.

UNITED NATIONS. **E-government Survey 2018:** Gearing e-Government to Support Transformation Towards Sustainable and Resilient Societies. New York: United Nations, 2018.

VASQUES, D. G. *et al.* Mineração de textos para gestão de clientes em empresas de telecomunicações. *In:* BRAZILIAN SYMPOSIUM ON INFORMATION SYSTEMS. 13., 2017, Lavras. **Anais** [...]. Lavras: Universidade Federal de Lavras, 2017. p. 9-16. Disponível em: https://sol.sbc.org.br/index.php/sbsi/article/view/6020/5918. Acesso em: 9 set. 2020.

VIGODA, E. From responsiveness to collaboration: governance, citizens, and the next generation of public administration. **Public Administration Review,** v. 62, n. 5, p. 527-540, 2002. Disponível em: https://www.researchgate.net/publication/227547066_From_Responsiveness_to_Collaboration_Governance_Citizens_and_the_Next_Generation_of_Public_Administration. Acesso em: 10 jul. 2020.

VIJAYARANI, S.; MUTHULAKSHMI, M. Comparative Analysis of Bayes and Lazy Classification Algorithms. **International Journal of Advanced Research in Computer and Communication Engineering,** v. 2, n. 8, p. 3118-3124, 2013.

WANG, H. Nearest neighbors by neighborhood counting. **IEEE Transactions on Pattern Analysis And Machine Intelligence,** v. 28, n. 6, p. 942-953, 2006.

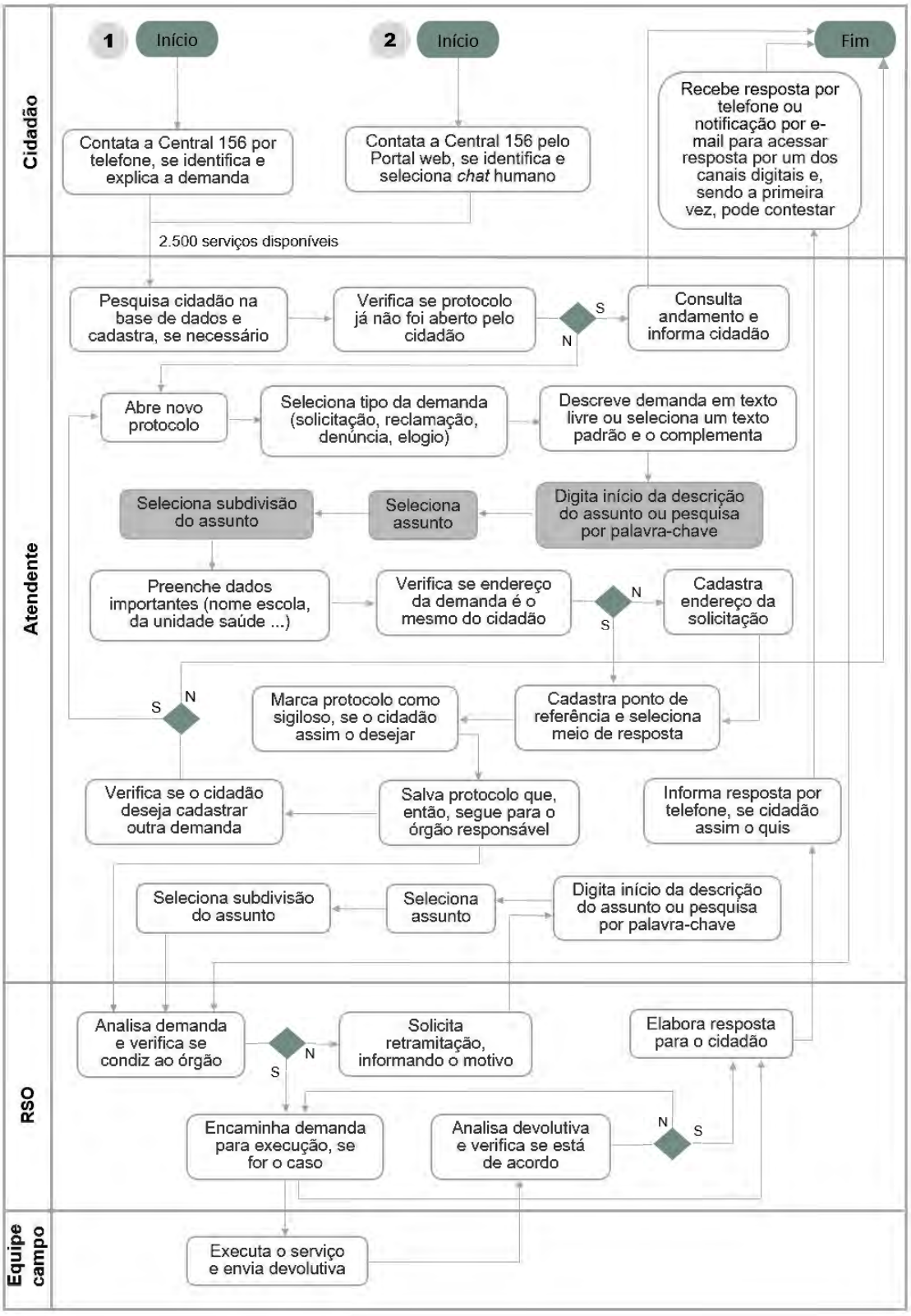
WEISS, S. M.; INDURKHYA, N.; ZHANG, T. **Fundamentals of Predictive Text Mining.** London: Springer, 2015.

WITTEN, I. H. **Text mining.** 2004. Disponível em: https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiqpbD8iMbzAhXiqpUCHemQDEgQFnoECAQQAQ&url=https%3A%2F%2Fwww.cms.waikato.ac.nz%2F~ihw%2Fpapers%2F04-IHW-Textmining.pdf&usg=AOvVaw3LbNlxqBQurWIl7qPsDGAv. Acesso em: 20 jun. 2019.

WITTEN, I. H.; FRANK, E.; HALL, M. A.; PAL, C. J. **Data mining:** Practical Machine Learning Tools and Techniques. 4th ed. Cambridge: Morgan Kaufmann, 2017.

WU, W. Citizen relationship management system users' contact channel choices: digital approach or call approach? **Information,** v. 8, n. 1, p. 1-16, 2017. Disponível em: https://www.mdpi.com/2078-2489/8/1/8/htm. Acesso em: 20 maio 2019.

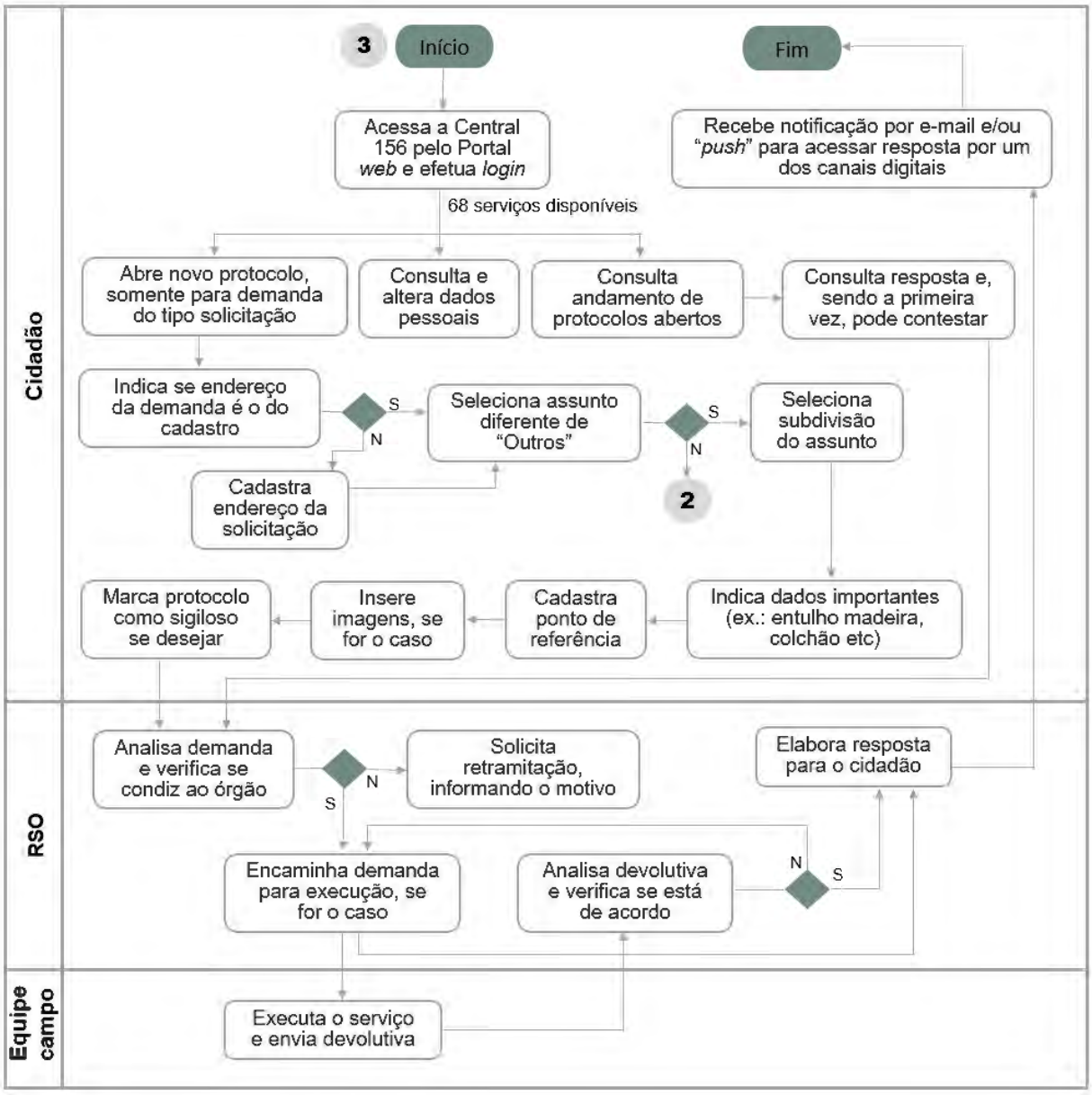
APÊNDICE 1 – FLUXO SIMPLIFICADO DE ATENDIMENTO AO CIDADÃO POR MEIO DOS CANAIS TELEFONE E *CHAT* HUMANO



Etapas envolvidas na pesquisa

FONTE: A autora (2021) com base em ICI (2021d)

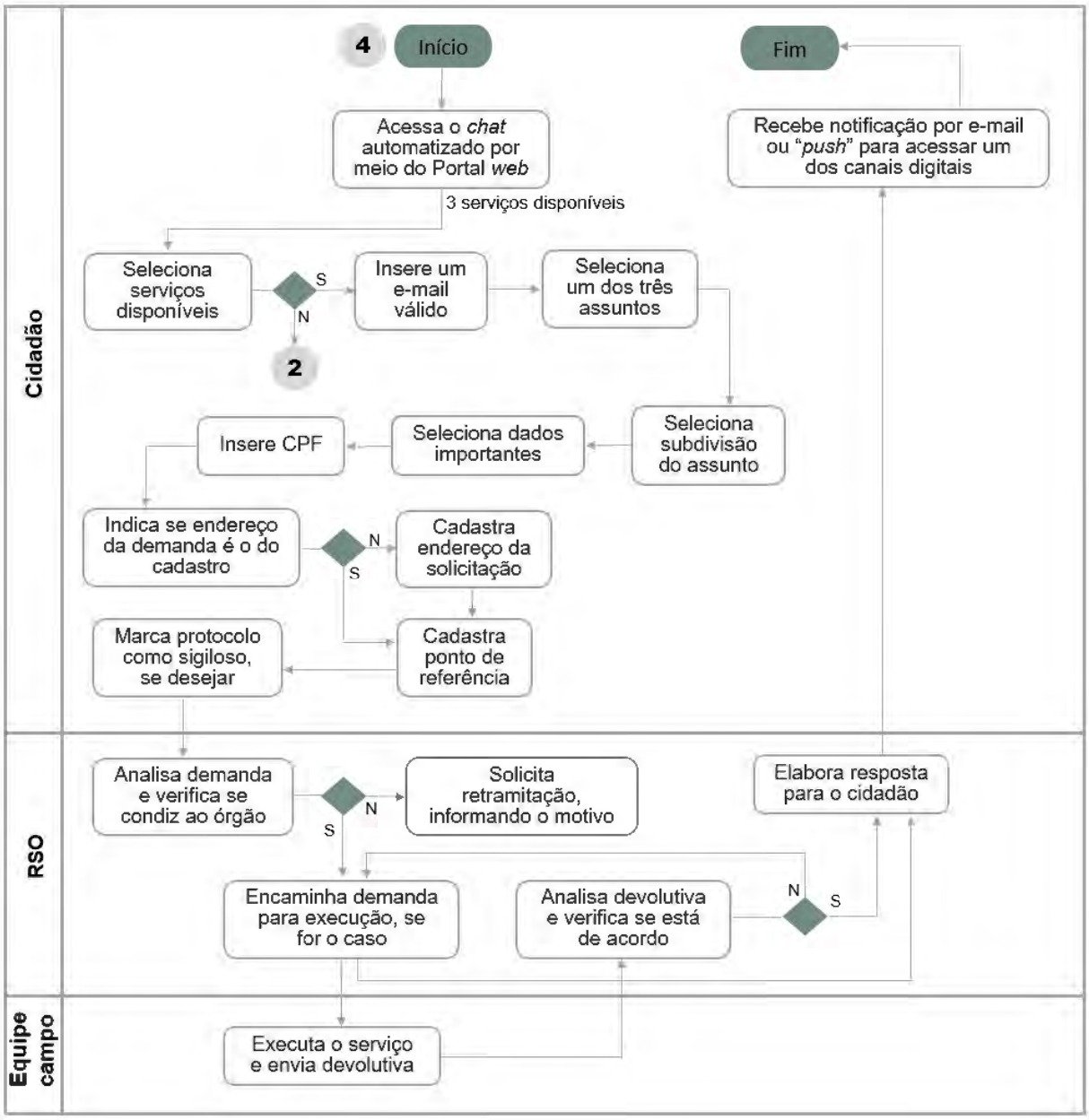
APÊNDICE 2 – FLUXO SIMPLIFICADO DE ATENDIMENTO AO CIDADÃO POR MEIO DO CANAL PORTAL *WEB*



FONTE: A autora (2021) com base em ICI (2021d)

Os serviços disponíveis referem-se aos assuntos: academia ao ar livre, acessibilidade, alvará, animais, árvore, bueiros, calçadas, coleta, coronavírus (Covid-19), Disque Solidariedade, dengue, fiscalização de comércios, fiscalização de obra particular, iluminação pública, imposto sobre serviço (ISS), limpeza, lombada física, outros, pavimentação, praças, proteção ao patrimônio, semáforo, sinalização, situação de rua – adulto / idoso, terreno baldio e trânsito.

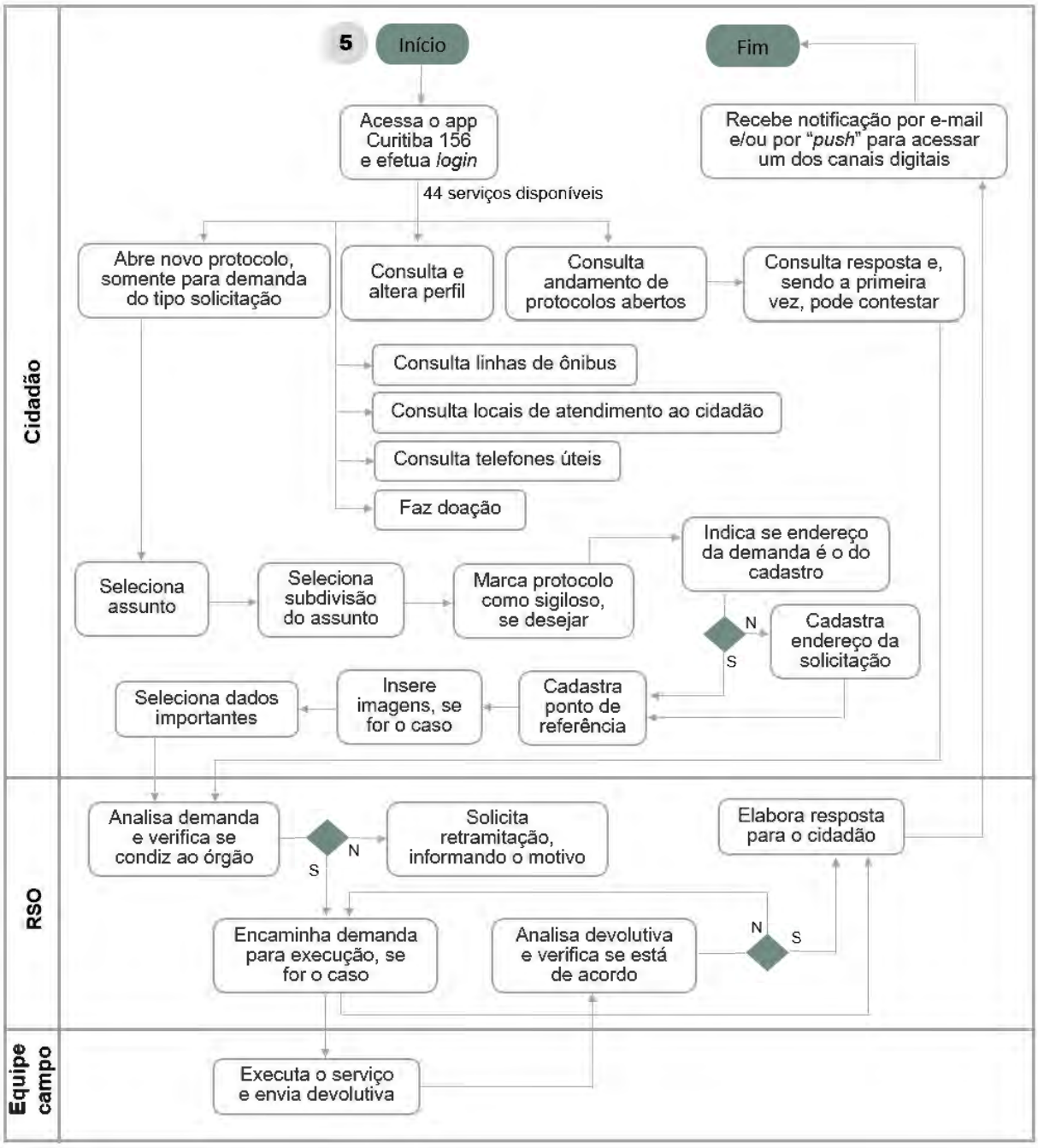
APÊNDICE 3 – FLUXO SIMPLIFICADO DE ATENDIMENTO AO CIDADÃO POR MEIO DO CANAL *CHAT* AUTOMATIZADO



FONTE: A autora (2021) com base em ICI (2021d)

Os serviços disponíveis referem-se aos assuntos pavimentação, coleta de resíduos vegetais e iluminação pública.

APÊNDICE 4 – FLUXO SIMPLIFICADO DE ATENDIMENTO AO CIDADÃO POR MEIO DO CANAL APP CURITIBA 156 - *MOBILE*



FONTE: A autora (2021) com base em ICI (2021e)

Os serviços disponíveis referem-se aos assuntos academia ao ar livre, acessibilidade, árvore, calçadas, coleta, coronavírus (Covid-19), dengue, fiscalização de obra particular, iluminação pública, lombada física, pavimentação, proteção ao

patrimônio, semáforo, situação de rua – adulto / idoso, terreno baldio, trânsito e violência contra criança e adolescente.

Sempre que o protocolo é sigiloso, o responsável no órgão não possui acesso aos dados do cidadão.

APÊNDICE 5 – CÁLCULO PARA OBTENÇÃO DOS EXEMPLOS, UTILIZANDO *UNDERSAMPLING* COM REGISTROS DA SMMA

Serviço	Total demandas	Valor utilizado	Valor amostral mínimo*
Coleta - resíduos vegetais de jardim	32.729	425	380
Coleta - entulhos diversos (pequena quantidade)	13.522	425	374
Coleta - caliças até 5 carrinhos de mão	9.838	425	370
Árvore - atendimento emergencial	1.981	425	322
Poluição - sonora - noturna	1.706	425	314
Poluição - sonora - diurna	1.579	425	310
Animais domésticos - remoção de caninos mortos	1.420	425	303
Poluição - atmosférica	1.414	425	303
Coleta - outros	1.107	425	286
Árvore - poda de manutenção em via pública	938	425	273
Animais - maus tratos - animais domésticos	913	425	271
Animais domésticos - cães e gatos acidentados em vias e logradouros públicos	892	425	269
Árvore - corte irregular - desmate	880	425	268
Animais domésticos - remoção de felinos mortos	587	425	233
Poluição - residual	573	425	231
Sinalização vertical - implantação de proibitiva para lixo	462	425	210
Animais - cães bravos ou avançando em via pública - fiscalização	433	425	204
Árvore - avaliação técnica - via pública	425	425	203
Coleta - atendimento (domiciliar)	323	323	--
Fiscalização de saneamento - vazamento de esgoto	298	298	--
Limpeza - óleo na pista	296	296	--
Fiscalização - resíduos dispostos inadequadamente em passeio	287	287	--
Limpeza - lavagem de vias - caminhão pipa	244	244	--
Varrição - manual	242	242	--
Limpeza - roçada em vias de competência do malp	242	242	--
Coleta - restos de podas da pmc	234	234	--
Árvore - adote uma árvore	229	229	--
Coleta - atendimento (resíduos recicláveis)	203	203	--
Árvore - sugestão de plantio	193	193	--
Animais domésticos - apreensão animais de grande porte soltos em via pública	190	190	--
Fiscalização de saneamento - lançamento de esgoto em via pública	157	157	--
Parques - conservação	141	141	--
Fiscalização de saneamento - lançamento de esgoto a céu aberto	134	134	--
Fiscalização de saneamento - lançamento de esgoto em corpos hídricos	134	134	--
Fiscalização de saneamento - mau cheiro na galeria de águas pluviais	125	125	--
Poluição - hídrica - empresas	110	110	--
Árvore - avaliação técnica - outros logradouros públicos	73	73	--
Animais domésticos - remoção de animais mortos - diagnóstico de raiva e febre amarela	72	72	--

Praças - conservação	69	69	--
Árvore - preventiva	68	68	--
Parques - manutenção	57	57	--
Limpeza - roçada em terrenos públicos	52	52	--
Bosques - conservação	51	51	--
Animais - abelhas	47	47	--
Jardim botânico - conservação	35	Serviços desconsiderados - possuem menos de 36 registros	
Animais - maus tratos - animais silvestres	34		
Animais - maus tratos - comércio de animais de estimação	28		
Praças - implantação de equipamentos	28		
Praças - manutenção	27		
Fiscalização de saneamento - canalização ou intervenção em curso de água	26		
Varrição - mecânica em vias com meio-fio	26		
Zoológico e passeio público - situações gerais	25		
Animais domésticos - informações animais: doenças, vacinação, mordidas	24		
Cemitérios municipais - limpeza	21		
Floreiras - substituição de vasos de flores em logradouros públicos	20		
Parques - implantação de equipamentos	17		
Árvore - corte de tocos / remoção de tocos / destoca	16		
Bosques - manutenção	16		
Coleta - atendimento (lixo tóxico)	15		
Coleta - implantação (resíduos recicláveis)	8		
Jardinetes - conservação	8		
Cemitérios municipais - reforma e revitalização	8		
Serviço funerário municipal - serviço mal prestado por funerária	8		
Praças - reforma e revitalização	8		
Animais domésticos - remoção de equídeos mortos	7		
Árvore - doação de mudas de árvores a pmc	6		
Bancos em ruas - implantação, recuperação e manutenção	6		
Cemitérios municipais - capelas	6		
Bosques - paisagismo - manejo da arborização	5		
Praças - implantação de área de lazer	5		
Parques - manutenção de cercas, muros e vedações de terrenos	5		
Serviço funerário municipal - fiscalização de funerárias e hospitais	4		
Fiscalização - lixo tóxico	4		
Cemitérios municipais - segurança	4		
Parques - reforma e revitalização	4		
Câmbio verde - atendimento ao público	4		
Coleta - implantação (domiciliar)	3		
Parques - paisagismo - manejo da arborização	3		
Parques - implantação de área de lazer	3		
Cemitérios municipais - iluminação	3		
Parques - paisagismo - plantio	3		
Gabinete - smma	3		
Serviço funerário municipal - solicitação de documentos	3		
Animais domésticos - remoção de bovídeos mortos	3		
Bosques - implantação de equipamentos	2		
Câmbio verde - implantação	2		

Cemitérios municipais - implantação de calçada e anti-pó nas vias internas	2	Serviços desconsiderados - possuem menos de 36 registros
Eixos de animação - conservação	2	
Suporte de floreiras em ruas - manutenção e recuperação	2	
Parques - fiscalização de comércio	2	
Jardinetes - manutenção	1	
Animais - apreensão de cães soltos	1	
Serviço funerário municipal - informações sobre plano de luto	1	
Praças - paisagismo - manejo da arborização	1	
Museu de história natural - situações gerais	1	
Eixos de animação - paisagismo - manejo da arborização	1	
Jardinetes - reforma e revitalização	1	
Bosques - fiscalização de comércio	1	
Jardinetes - implantação de área de lazer	1	
Câmbio verde - produtos	1	
Largos - paisagismo - manejo da arborização	1	
Total	76.175	

* Amostragem considerando-se 95% de grau de confiança e 5% de margem de erro.

FONTE: A autora (2021) com base em PMC (2019a)

APÊNDICE 6 – CÁLCULO PARA OBTENÇÃO DOS EXEMPLOS, UTILIZANDO *UNDERSAMPLING* PARA AS DEMAIS CLASSES DO ESTUDO

Classe	Serviço	Total demandas	Valor utilizado	Valor amostral mínimo*
FAS	Abordagem social de rua - pessoas/famílias em desabrigo na rua	11.376	1.452	372
FAS	Abordagem social de rua - dormindo/caídas na rua	6.772	1.452	364
FAS	Disque solidariedade - doações	3.911	1.452	350
FAS	Abordagem social de rua - perdida/desorientada	1.452	1.452	304
FAS	Abordagem social de rua - alcoolizadas/drogadas	1.088	1.088	--
FAS	Abordagem social de rua - esmolando	773	773	--
FAS	Abordagem social de rua - adulto desaparecido	614	614	--
FAS	Abordagem social de rua - operação inverno	554	554	--
FAS	Abordagem social de rua - crianças desenvolvendo atividade informal	313	313	--
FAS	Funcionários - FAS	102	102	--
FAS	Gabinete - FAS	54	54	--
SMDT	Trânsito - fiscalização de estacionamento irregular	39.367	452	381
SMDT	Trânsito - fiscalização imediata	2.791	452	338
SMDT	Trânsito - fiscalização programada	2.318	452	330
SMDT	Semáforo - lâmpada apagada	1.744	452	315
SMDT	Trânsito - fiscalização de bloqueio de pista	1.500	452	306
SMDT	Trânsito - veículo abandonado	1.476	452	305
SMDT	Fiscalização - cabos danificados	1.158	452	289
SMDT	Semáforo - em alerta	1.039	452	281
SMDT	Semáforo - manutenção	843	452	265
SMDT	Passeio - danos causados por obras da: Sanepar, Copel, Brasil Telecom e outros	842	452	265
SMDT	Lombada física - implantação	497	452	217
SMDT	Fiscalização - acompanhamento de obras de concessionárias em via pública	466	452	211
SMDT	Segurança de edificações e imóveis - estruturas	452	452	--
SMDT	Proteção ao cidadão - GM - solicitação de maior segurança	396	396	--
SMDT	Trânsito - agente de trânsito	340	340	--
SMDT	Sinalização de estacionamento - proibição de estacionamento	327	327	--
SMDT	Proteção ao cidadão - GM - solicitação de guardas municipais	297	297	--
SMDT	Semáforo - implantação	269	269	--
SMDT	Sinalização horizontal - repintura	263	263	--
SMDT	SMDS - reclamações	239	239	--
SMDT	Semáforo - falha de sincronismo	238	238	--
SMDT	Proteção ao patrimônio - GM - invasão de equipamento edificado	225	225	--
SMDT	SMDS - elogios	218	218	--
SMDT	Proteção ao cidadão - GM - fornecimento de lona	204	204	--
SMDT	Semáforo - tempos de verde	192	192	--
SMDT	Pavimento danificado - danos causados por obras da Sanepar, Copel, Brasil Telecom e outros	190	190	--
SMDT	Infrações de trânsito - outras informações	179	179	--

SMDT	Sentido de tráfego - alteração de sentido de tráfego	178	178	--
SMDT	Faixa de travessia elevada - implantação	124	124	--
SMDT	Estar - agentes do estar	113	113	--
SMDT	Sinalização vertical - reposição de placas	74	74	--
SMDT	Proteção ao patrimônio - GM - violação de alarme	72	72	--
SMDT	Acessibilidade - implantação de rampa	70	70	--
SMDT	Velocidade na via - excesso de velocidade	68	68	--
SMDT	Sinalização horizontal - pintura de legenda: pare	64	64	--
SMDT	Infrações de trânsito - processos de defesa prévia, Jari e Cetran	62	62	--
SMDT	Sinalização vertical - manutenção de placas	61	61	--
SMDT	Radar de velocidade - implantação	55	55	--
SMDT	Estar - implantação	55	55	--
SMDT	Sinalização horizontal - colocação de tartarugas	54	54	--
SMDT	Sinalização horizontal - pintura de faixas divisórias de pista	53	53	--
SMDT	Proteção ao patrimônio - GM - invasão do transporte coletivo	53	53	--
SMDT	Sinalização vertical - proibição de tráfego/estacionamento de caminhões	47	47	--
SMDT	Correção geométrica - implantação	47	47	--
SMDT	Animais - dejetos gerados pelos animais de estimação	45	45	--
SMDT	Estar - outros	43	43	--
SMDT	Sinalização horizontal - repintura de lombada	39	39	--
SMDT	Proteção ao cidadão - GM - sugestão de segurança	37	37	--
SMOP	Iluminação pública - via pública - manutenção de luminárias em vias públicas	26.952	4.000	379
SMOP	Iluminação pública - manutenção de luminárias	8.281	4.000	368
SMOP	Iluminação pública em logradouros públicos - manutenção de iluminação em logradouros públicos	633	633	--
SMOP	Iluminação pública - via pública - melhoria de iluminação em vias públicas	392	392	--
SMOP	Iluminação pública - via pública - implantação de iluminação em vias públicas	244	244	--
SMOP	Iluminação pública - via pública - implantação de extensão/ampliação de rede em vias públicas	228	228	--
SMOP	Fiscalização - pavimentação definitiva ou alternativa (obra em andamento)	201	201	--
SMOP	Praças - manutenção de iluminação	157	157	--
SMOP	Iluminação pública - implantação de aumento de potência	141	141	--
SMOP	Iluminação pública - implantação de luminárias	118	118	--
SMOP	Drenagem - matriz - manutenção em galeria de águas pluviais	108	108	--
SMOP	Drenagem - matriz - limpeza / manutenção de caixa de captação	102	102	--
SMOP	Iluminação pública - implantação de extensão de rede	91	91	--
SMOP	Iluminação pública - via pública - contribuição (taxa)	80	80	--
SMOP	Fiscalização - fresagem e recape	59	59	--
SMOP	Pontes de madeira - ponte danificada (permitindo passagem)	55	55	--
SMOP	Macrodrenagem - desobstrução dos cursos hídricos - rios e córregos	54	54	--
SMOP	Drenagem - matriz - desobstrução em galeria de águas pluviais	51	51	--

SMOP	Macrodrenagem - erosão em margem de rio ou córrego	47	47	--
SMOP	Macrodrenagem - desassoreamento de rios, córregos e canais	47	47	--
SMOP	Drenagem - matriz - reposição de grelha	47	47	--
SMOP	Iluminação pública - contribuição (taxa)	45	45	--
SMOP	Drenagem - matriz - reposição de tampão	42	42	--
SMOP	Drenagem - matriz - manutenção em caixa de captação	42	42	--
SMOP	Fiscalização de terrenos baldios ou edificados - vedação de terreno e imóvel edificado do município	38	38	--
SMS	Posto de saúde - recursos humanos - atendimento profissional	4.498	340	355
SMS	Fauna sinantrópica - risco para leptospirose/ roedores em bueiro	2.571	340	335
SMS	Posto de saúde - fluxo de atendimento	2.133	340	326
SMS	Posto de saúde - agendamento de consulta especializada / SADT	1.646	340	312
SMS	Vigilância em saúde ambiental - inspeção/orientação - mosquito da dengue	871	340	267
SMS	Unidade de saúde 24h - recursos humanos - atendimento profissional	843	340	265
SMS	Vigilância sanitária em estabelecimentos - fiscalização em serviços de interesse a saúde	820	340	262
SMS	Posto de saúde - agendamento de consulta básica	799	340	260
SMS	Posto de saúde - recursos humanos - quantidade de profissionais	600	340	235
SMS	Profissionais/serviços credenciados - fluxo de atendimento	577	340	231
SMS	Unidade de saúde 24h - recursos humanos - atendimento médico/outros profissionais FEAES/OS	568	340	230
SMS	Fauna sinantrópica - orientação, manejo e controle de doenças e agravos	547	340	226
SMS	Orientações sanitárias para residências - lixo/água e outras irregularidades	539	340	225
SMS	Posto de saúde - aplicativo Saúde Já Curitiba	512	340	220
SMS	Fauna sinantrópica - morcegos	487	340	216
SMS	Unidade de saúde 24h - demora no atendimento para consulta médica/outros profissionais feaes/os	476	340	213
SMS	Orientações sanitárias para residências - criação de animais domésticos	453	340	209
SMS	Posto de saúde - recursos materiais - linha telefônica	426	340	203
SMS	Posto de saúde - demora no atendimento	420	340	201
SMS	Posto de saúde - recursos materiais - medicamentos	417	340	201
SMS	Posto de saúde - recursos materiais - material médico hospitalar	340	340	181
SMS	Outros setores SMS - exames/ procedimentos de alto custo	273	273	--
SMS	Diretoria de urgência e emergência - registro de ocorrência	273	273	--
SMS	Posto de saúde - recursos humanos - recusa de atendimento	272	272	--
SMS	Posto de saúde - recursos humanos - falta do profissional ao trabalho	245	245	--
SMS	Unidade de saúde 24h - fluxo de atendimento	244	244	--
SMS	Diretoria de urgência e emergência - demora para internamento	219	219	--

SMS	Posto de saúde - recursos materiais - vacinas	202	202	--
SMS	Posto de saúde - recursos materiais - material permanente	199	199	--
SMS	Posto de saúde - assistência à gestante	168	168	--
SMS	Diretoria de urgência e emergência - recursos humanos - atendimento profissional	145	145	--
SMS	Centro de atenção psicossocial - caps - recursos humanos - atendimento profissional	140	140	--
SMS	Rede hospitalar/feaes - recursos humanos - atendimento profissional	134	134	--
SMS	Profissionais/serviços credenciados - recursos humanos - atendimento profissional	131	131	--
SMS	Unidade de saúde 24h - demora no atendimento	129	129	--
SMS	Vigilância sanitária de produtos - produtos de interesse a saúde	97	97	--
SMS	Posto de saúde - outros	93	93	--
SMS	Unidade de saúde 24h - recursos humanos - atendimento médico/outros profissionais feas/os	92	92	--
SMS	Profissionais/serviços credenciados - agendamento de consulta de retorno	92	92	--
SMS	Unidade de saúde 24h - recursos humanos - quantidade de médicos/outros profissionais feaes/os	84	84	--
SMS	Unidade de saúde 24h - demora no atendimento para consulta médica/outros profissionais feas/os	81	81	--
SMS	Posto de saúde - recursos materiais - outros	81	81	--
SMS	Rede hospitalar/feaes - fluxo de atendimento	80	80	--
SMS	Outros setores SMS - fluxo de atendimento	74	74	--
SMS	Posto de saúde - exame de coleta na unidade de saúde	72	72	--
SMS	Profissionais/serviços credenciados - demora no atendimento	64	64	--
SMS	Unidade de saúde 24h - recursos materiais - outros	63	63	--
SMS	Profissionais/serviços credenciados - recursos materiais - órtese e prótese	61	61	--
SMS	Outros setores SMS - recursos humanos - atendimento profissional	57	57	--
SMS	Outros setores SMS - recursos materiais - medicamentos	56	56	--
SMS	Unidade de saúde 24h - recursos materiais - material permanente	53	53	--
SMS	Posto de saúde - atenção nutricional - leites e dietas	52	52	--
SMS	Outros setores SMS - recursos materiais - vacinas	51	51	--
SMS	Fauna sinantrópica - orientação sobre roedores e leptospirose	49	49	--
SMS	Vigilância sanitária em hospitais, clínicas e Consultórios - fiscalização em hospitais, clínicas e consultórios	49	49	--
SMS	Outros setores sms - aplicativo saúde já curitiba	48	48	--
SMS	Posto de saúde - recursos humanos - alteração de profissional	47	47	--
SMS	Outros setores SMS - recursos materiais - material médico hospitalar	43	43	--
SMS	Outros setores SMS - recursos materiais - material permanente	42	42	--
SMS	Outros setores SMS - recursos materiais - manutenção e obras	42	42	--
SMS	Diretoria de urgência e emergência - outros	41	41	--

SMS	Outros setores SMS - outros	39	39	--
SMS	Profissionais/serviços credenciados - recursos materiais - linha telefônica	38	38	--
SMS	Diretoria de urgência e emergência - demora no atendimento	38	38	--
SMS	Posto de saúde - recursos humanos - outros	37	37	--
SMS	Outros setores SMS - transporte sanitário	37	37	--
SGM	Pavimentação - manutenção	4.711	1.170	356
SGM	Pavimentação - revitalização	2.820	1.170	339
SGM	Pavimentação - implantação	2.466	1.170	333
SGM	Drenagem - limpeza e desobstrução de caixa de captação	1.170	1.170	320
SGM	Drenagem - erosão em galeria de águas pluviais	918	918	--
SGM	Limpeza - limpeza e roçada de sarjeta - meio-fio	601	601	--
SGM	Limpeza - roçada em equipamento público	496	496	--
SGM	Passeio - recuperação de calçada	456	456	--
SGM	Limpeza - roçada de rua	439	439	--
SGM	Drenagem - reposição de tampa/ralo/grelha de caixa de captação	433	433	--
SGM	Atendimento - agradecimento à central	428	428	--
SGM	Drenagem - implantação de caixa de captação	253	253	--
SGM	Portal 156 - unificação de cadastro	190	190	--
SGM	Passeio - roçada	186	186	--
SGM	Drenagem - limpeza de tubulação até 80cm	151	151	--
SGM	Drenagem - recuperação de caixa de captação	151	151	--
SGM	Drenagem - sondagem em galerias de águas pluviais	145	145	--
SGM	Pavimentação - complementação até o meio-fio	142	142	--
SGM	Rua da cidadania - administração regional	122	122	--
SGM	Drenagem - recuperação de tubulação até 80cm	94	94	--
SGM	Drenagem - relocação de caixa de captação	87	87	--
SGM	Atendimento - reclamação da central	84	84	--
SGM	Aplicativo Curitiba 156 - informações gerais	83	83	--
SGM	Arruamento - nivelamento de rua	83	83	--
SGM	Drenagem - alagamentos	72	72	--
SGM	Prefeito - elogio	67	67	--
SGM	Pavimentação alternativa - antipó - revitalização	64	64	--
SGM	Passeio - recuperação de meio-fio	60	60	--
SGM	Drenagem - assentamento de tubulação até 80cm	54	54	--
SGM	Portal 156 - login de acesso	51	51	--
SGM	Pavimentação alternativa - antipó - tapa buraco	48	48	--
SGM	Drenagem - limpeza de valeta	45	45	--
SGM	Pavimentação alternativa - antipó - implantação	44	44	--
SGM	Prefeito - solicitações	38	38	--
URBS	Motoristas, cobradores e porteiros - atrasar o horário durante a operação	1.060	1.060	283
URBS	Cartão transporte - venda	765	765	256
URBS	Motoristas, cobradores e porteiros - supressões de horário	700	700	249
URBS	Motoristas, cobradores e porteiros - recusar passageiros sem motivo justificado	687	687	247
URBS	Cartão transporte - geral	549	549	227
URBS	Cartão transporte - integração temporal	478	478	214
URBS	Motoristas, cobradores e porteiros - adiantar o horário durante a operação	392	392	195

URBS	Ônibus - manutenção	367	367	188
URBS	Programação do transporte coletivo - melhoria na oferta de veículos/viagens	365	365	188
URBS	Motoristas, cobradores e porteiros - tratar passageiros com falta de urbanidade	334	334	188
URBS	Motoristas, cobradores e porteiros - dirigir inadequadamente com risco de acidentes	280	280	188
URBS	Motoristas, cobradores e porteiros - deixar de manter atitudes condizentes com a função	234	234	188
URBS	Motoristas, cobradores e porteiros - elogio pelo bom desempenho da função	225	225	188
URBS	Passe escolar - geral	221	221	188
URBS	Transporte coletivo - reclamação contra a fiscalização do transporte	181	181	--
URBS	Internet - site - Urbs	177	177	--
URBS	Motoristas, cobradores e porteiros - solicitação de fiscalização	172	172	--
URBS	Motoristas, cobradores e porteiros - partir com ônibus com passageiro embarcando / desembarcando	171	171	--
URBS	Motoristas, cobradores e porteiros - deixar de cumprir determinação da Urbs	163	163	--
URBS	Transporte comercial (fretamento) - veículos clandestinos	158	158	--
URBS	Motoristas, cobradores e porteiros - dirigir inadequadamente desobedecendo às regras de trânsito	139	139	--
URBS	Táxi - outros	137	137	--
URBS	Táxi - esquecimento	137	137	--
URBS	Motoristas, cobradores e porteiros - vendedores ambulantes e/ou pedintes em ônibus, terminais e estações-tubo	128	128	--
URBS	Ponto de ônibus - manutenção	125	125	--
URBS	Itinerário do transporte coletivo - alteração	123	123	--
URBS	Ponto de ônibus - relocação	119	119	--
URBS	Ônibus - substituição por novos	115	115	--
URBS	Rodoviária - outros	114	114	--
URBS	Táxi - direção perigosa	93	93	--
URBS	Motoristas, cobradores e porteiros - denúncia/reclamação de desvio de itinerário	91	91	--
URBS	Motoristas, cobradores e porteiros - dirigir inadequadamente pondo em risco aos passageiros	89	89	--
URBS	Terminais/estações tubo (obras) - espaço físico nas estações e terminais	84	84	--
URBS	Ponto de ônibus - implantação	84	84	--
URBS	Programação do transporte coletivo - alteração de horário	81	81	--
URBS	Lixeiras em ruas - manutenção e reposição	78	78	--
URBS	Painéis de mensagens variáveis - implantação e manutenção	77	77	--
URBS	Novo mobiliário urbano - ponto de ônibus	73	73	--
URBS	Ônibus - dedetização	67	67	--
URBS	Ônibus - sugestão de melhoria	65	65	--
URBS	Motoristas, cobradores e porteiros - não atender ao sinal de parada para desembarque	64	64	--
URBS	Transporte escolar - direção perigosa	64	64	--
URBS	Cartão transporte - problemas no validador	63	63	--
URBS	Transporte escolar - veículo clandestino	63	63	--

URBS	Motoristas, cobradores e porteiros - alterar ponto de parada sem autorização	60	60	--
URBS	Ônibus - equipamentos do veículo	52	52	--
URBS	Táxi - agressão verbal ou física	51	51	--
URBS	Táxi - não tratar com polidez e urbanidade	48	48	--
URBS	Outros funcionários do transporte coletivo - vigilantes/segurança	44	44	--
URBS	Internet - help desk bilhetagem eletrônica	41	41	--
URBS	Programação do transporte coletivo - criação de nova linha	41	41	--
URBS	Manutenção predial - terminais de ônibus	40	40	--
URBS	Cabine de integração - ruas da cidadania ou shopping popular	39	39	--
URBS	Terminais/estações tubo (projetos) - espaço físico nas estações e terminais	39	39	--
URBS	Ponto de ônibus - implantação de abrigo	37	37	--
URBS	Espaço comercial - ocupação comercial em espaços públicos	37	37	--
URBS	Estação tubo - porta travada fechada	36	36	--
URBS	Estação tubo - objetos perdidos nas estações tubo	36	36	--
SMU	Fiscalização de terrenos baldios ou edificados - limpeza	3.101	3.101	--
SMU	Passeio - obstrução do passeio	1.761	1.761	--
SMU	Fiscalização de obras - alvará de construção	1.507	1.507	--
SMU	Passeio - fiscalização de construção/reconstrução de passeio	1.231	1.231	--
SMU	Fiscalização do comércio estabelecido - comércio diurno	915	915	--
SMU	Fiscalização de publicidade - publicidade em via pública (banners, cavaletes e faixas)	635	635	--
SMU	Fiscalização de terrenos baldios ou edificados - construção/reconstrução de muros	283	283	--
SMU	Placas de nomenclatura de ruas - manutenção	256	256	--
SMU	Placas de nomenclatura de ruas - colocação em toda a extensão	194	194	--
SMU	Fiscalização de mocós - utilização ilegal de imóvel por população de rua (que gere insegurança)	178	178	--
SMU	Fiscalização do comércio estabelecido - comércio noturno	145	145	--
SMU	Numeração predial - correção total da rua	92	92	--
SMU	Fiscalização do comércio ambulante - comércio de produtos não licenciados em via pública	71	71	--
SMU	Fiscalização de terrenos baldios ou edificados - erguer guia rebaixada	57	57	--
SMU	Fiscalização de obras - alvará de demolição	43	43	--
SMU	Numeração predial - correção de dupla na rua	41	41	--
SMU	Fiscalização do comércio ambulante - comércio de produtos fora do horário e local liberados	37	37	--

* Amostragem considerando-se 95% de grau de confiança e 5% de margem de erro.

FONTE: A autora (2021) com base em PMC (2019a)

APÊNDICE 7 – LISTA DE STOPWORDS UTILIZADAS

a	dois	for	jaqueta	quarenta	temos
anexa	dos	fora	jeans	quatorze	tenha
anexo	doze	foram	lhe	quatro	tenho
ao	duzentos	fossemos	lhes	quatrocentos	tenis
aos	e	formos	link	que	tera
aquela	ela	fosse	mais	quem	terá
aquelas	elas	fossem	mas	quinhentos	terao
aquele	ele	foramos	me	quinze	terão
aqueles	eles	fôramos	mesmo	refere	terei
aquilo	em	forem	meu	referente	teremos
as	entre	fôssemos	meus	relata	teria
às	era	foto	mil	relatada	teriam
até	eram	fui	milhao	saber	teriamos
ate	eramos	gostaria	minha	sao	teríamos
atraves	éramos	ha	minhas	são	teu
azul	essa	há	moletom	se	teus
blusa	essas	haja	moleton	segue	teve
branca	esse	hajam	muito	seguir	tinha
branco	esses	hajamos	na	seis	tinham
calca	esta	hao	nao	seiscentos	tinhamos
camiseta	está	hão	não	seja	tínhamos
casaco	estamos	havemos	nas	sejam	tinhas
cem	estao	hei	nem	sejamos	tive
chat	estão	hoje	no	sem	tivemos
cidada	estas	houve	nos	sera	tiver
cidadao	estava	houvemos	nós	será	tivera
cinco	estavam	houver	nossa	serao	tiveram
cinquenta	estavamos	houvera	nossas	serão	tiveramos
cinza	estávamos	houverá	nosso	serei	tivéramos
cita	este	houvéram	nossos	seremos	tiverem
citada	esteja	houveram	nove	seria	tivermos
citado	estejam	houveramos	novecentos	seriam	tivesse
com	estejamos	houvéramos	noventa	seriamos	tivessem
como	estes	houverao	num	seríamos	tivessemos
da	esteve	houverão	numa	sessenta	tivéssemos
das	estive	houverei	oitenta	sete	trata
data	estivemos	houverem	oito	setecentos	três
de	estiver	houveremos	oitocentos	setenta	treze
dela	estivera	houveria	onze	seu	trezentos
delas	estiveram	houveriam	os	seus	trinta
dele	estiveramos	houveriamos	ou	situacao	tu
deles	estivéramos	houveríamos	para	só	tua
depois	estiverem	houvesse	pela	so	tuas
descricao	estivermos	houvessem	pelas	solicita	um
dez	estivesse	houvessemos	pelo	somos	uma
dezenove	estivessem	houvéssemos	pelos	sou	vermelha
dezesseis	estivessemos	informa	perante	sua	vermelho
dezessete	estivéssemos	integra	por	suas	vinte
dezoito	estou	internet	preto	tambem	voce
dia	eu	isso	protocolo	também	você
dias	foi	isto	qual	te	voces
do	fomos	ja	quanto	tem	vocês

Palavras na cor azul acrescidas pela autora

FONTE: A autora (2021) com base no pacote “tm” do software R

APÊNDICE 8 – DETALHAMENTO DO PROCESSO DE GESTÃO DA INFORMAÇÃO DA PMC, REALIZADO VIA CENTRAL 156

Etapa	Central 156 - Tarefas
Etapa 1 - Identificação das necessidades de informação	● As necessidades, compreendendo informações, fluxos de tarefas e funcionalidades, são estabelecidos e reavaliados, de tempos em tempos, pelo órgão gestor do 156, a SGM e pelo prestador de serviços, o ICI; ● Todos os órgãos são consultados e as necessidades registradas no documento SGM - Evolução SIAC 156.
Etapa 2 - Coleta de informação	● A coleta é realizada por meio dos cinco canais disponibilizados à população: o número telefônico 156, o portal *web*, o *chat* humano, o *chat* automatizado, ou robô, e o aplicativo *mobil*e Curitiba 156; ● Os dados dos cidadãos são coletados a partir do registro pelo próprio cidadão, ou por um dos atendentes da Central 156 quando utilizado o canal telefone; ● Quando utilizados os canais telefone e *chat* humano, as demandas em texto livre são coletadas pelos atendentes da Central 156 e registradas no SIAC, no campo “descrição”; ● As demandas registradas pelos demais canais são selecionadas pelo cidadão, com limitação de serviços ofertados; ● As demandas recebem número de protocolo sequencial e automático.
Etapas 3 e 4 - Classificação e tratamento de informação	● As demandas textuais são classificadas manualmente, segundo assunto e subdivisão, por um dos atendentes da Central 156: o caso o atendente saiba o assunto, digita o início de sua denominação e o SIAC apresenta o assunto; caso contrário, o atendente efetua a busca, digitando uma palavra ou termo chave e assim, o sistema carrega uma lista com assuntos que contêm a palavra ou o termo para seleção; o ao selecionar o assunto, as subdivisões específicas são carregadas na lista para seleção e podem ser pesquisadas da mesma forma que os assuntos; ● Há tratamento para minimizar o preenchimento incorreto dos dados; ● O armazenamento ocorre em banco de dados corporativo gerenciado pelo ICI.
Etapa 5 - Desenvolvimento de produtos e serviços de informação	• Os produtos e serviços podem ser compreendidos como o Sistema 156 de Atendimento ao Cidadão (SIAC), suas ferramentas e relatórios: o Operacionais: Protocolos criados; Protocolos respondidos no período; Demanda por período; Protocolos x situação x resposta x logradouro; o Autoridades: Protocolos de autoridade – proposições; Demandas de autoridade – pedidos de informação e projetos de lei; o Gerenciais: Gerenciais e Top 10; o Case BI-156, com histórico das demandas por órgão, período, assunto e subdivisão.

Etapa	Central 156 - Tarefas
Etapa 6 - Distribuição de informação	● Encaminhamento, via sistema e a partir da classificação efetuada, das demandas aos órgãos competentes para análise e providências; ● Comunicação da resposta para o cidadão, por e-mail ou por telefonema.
Etapa 7 - Análise e uso de informação	● Uso da informação pelo cidadão; ● Uso das demandas pela PMC para planejamento urbano e desenvolvimento de programas e ações de interesse comum; ● Consumo da informação por outros sistemas e aplicações, como o Sistema integrado de gerenciamento da manutenção urbana (SIGMU), Sistema de governo técnico legislativo (GTL), Sistema único de autoridades (SUA), Sistema de proposição legislativa (SPL), sistemas de iluminação pública, coleta de lixo e arborização e o Sistema da Central de Informações, que reúne informações estratégicas de todos os órgãos.

FONTE: A autora (2021), com base em McGee e Prusak (1994, p. 107-126) e SIAC-156 (ICI, 2021b)

APÊNDICE 9 – RESULTADOS OBTIDOS NA CLASSIFICAÇÃO COM UNIGRAMAS – *DATASET* MAIOR

uni999 - resultado J48 - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     uni999-3-weka.filters.unsupervised.attribute.Remove-R1
Instances:    37588
Attributes:   1902
              [list of attributes omitted]
Test mode:    10-fold cross-validation
```

uni999 - resultado J48 - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
Correctly Classified Instances       34238               91.0876 %
Incorrectly Classified Instances      3350                8.9124 %
Kappa statistic                          0.8981
Mean absolute error                      0.0304
Root mean squared error                  0.1399
Relative absolute error                 13.9253 %
Root relative squared error             42.3138 %
Total Number of Instances            37588

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,984    0,002    0,984      0,984    0,984      0,982    0,993     0,980     FAS
                 0,874    0,032    0,792      0,874    0,831      0,807    0,935     0,801     SGM
                 0,836    0,018    0,868      0,836    0,851      0,831    0,941     0,820     SMDT
                 0,877    0,014    0,907      0,877    0,892      0,875    0,954     0,884     SMMA
                 0,922    0,011    0,917      0,922    0,920      0,909    0,966     0,893     SMU
                 0,937    0,009    0,932      0,937    0,935      0,926    0,978     0,926     URBS
                 0,926    0,004    0,969      0,926    0,947      0,940    0,980     0,941     SMOP
                 0,939    0,010    0,933      0,939    0,936      0,926    0,980     0,936     SMS
Weighted Avg.    0,911    0,013    0,912      0,911    0,911      0,899    0,966     0,897
```

uni999 - resultado NBMultinomial - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayesMultinomial
Relation:     uni999-3-weka.filters.unsupervised.attribute.Remove-R1
Instances:    37588
Attributes:   1902
              [list of attributes omitted]
Test mode:    10-fold cross-validation

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances       34312               91.2845 %
Incorrectly Classified Instances      3276                8.7155 %
Kappa statistic                          0.9003
Mean absolute error                      0.0219
Root mean squared error                  0.1457
Relative absolute error                  9.9965 %
Root relative squared error             44.0806 %
Total Number of Instances            37588

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,968    0,002    0,988      0,968    0,978      0,975    0,995     0,986     FAS
                 0,898    0,030    0,808      0,898    0,851      0,830    0,972     0,812     SGM
                 0,831    0,015    0,888      0,831    0,858      0,840    0,964     0,877     SMDT
                 0,918    0,016    0,902      0,918    0,910      0,895    0,983     0,929     SMMA
                 0,932    0,020    0,865      0,932    0,897      0,883    0,984     0,907     SMU
                 0,937    0,005    0,960      0,937    0,948      0,941    0,993     0,968     URBS
                 0,922    0,004    0,970      0,922    0,945      0,938    0,984     0,950     SMOP
                 0,903    0,008    0,946      0,903    0,924      0,913    0,987     0,952     SMS
Weighted Avg.    0,913    0,013    0,915      0,913    0,913      0,901    0,983     0,922
```

APÊNDICE 10 – RESULTADOS OBTIDOS NA CLASSIFICAÇÃO COM UNIGRAMAS – *DATASET* MENOR

uni9975-3 - resultado J48 - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     uni9975-3-weka.filters.unsupervised.attribute.Remove-R1
Instances:    37588
Attributes:   1004
              [list of attributes omitted]
Test mode:    10-fold cross-validation
```

uni9975-3 - resultado J48 - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
Correctly Classified Instances       34217               91.0317 %
Incorrectly Classified Instances      3371                8.9683 %
Kappa statistic                          0.8975
Mean absolute error                      0.0305
Root mean squared error                  0.1399
Relative absolute error                 13.9707 %
Root relative squared error             42.3108 %
Total Number of Instances            37588

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,984    0,002    0,984      0,984    0,984      0,981    0,993     0,978     FAS
               0,876    0,032    0,794      0,876    0,833      0,810    0,941     0,808     SGM
               0,830    0,021    0,850      0,830    0,840      0,818    0,935     0,816     SMDT
               0,888    0,012    0,921      0,888    0,904      0,889    0,963     0,897     SMMA
               0,924    0,012    0,915      0,924    0,919      0,908    0,967     0,904     SMU
               0,936    0,009    0,933      0,936    0,935      0,925    0,978     0,929     URBS
               0,927    0,004    0,971      0,927    0,949      0,942    0,981     0,946     SMOP
               0,926    0,011    0,931      0,926    0,929      0,918    0,974     0,927     SMS
Weighted Avg.  0,910    0,013    0,912      0,910    0,911      0,898    0,966     0,900
```

uni9975-3 - resultado k-NN - IBk - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Relation:     uni9975-3-weka.filters.unsupervised.attribute.Remove-R1
Instances:    37588
Attributes:   1004
              [list of attributes omitted]
Test mode:    10-fold cross-validation
```

uni9975-3 - resultado k-NN - IBk - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
Correctly Classified Instances       31781               84.5509 %
Incorrectly Classified Instances      5807               15.4491 %
Kappa statistic                          0.8234
Mean absolute error                      0.04
Root mean squared error                  0.1953
Relative absolute error                 18.316  %
Root relative squared error             59.0653 %
Total Number of Instances            37588

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,964    0,005    0,960      0,964    0,962      0,957    0,986     0,945     FAS
               0,833    0,043    0,731      0,833    0,779      0,748    0,910     0,633     SGM
               0,735    0,026    0,801      0,735    0,767      0,736    0,886     0,693     SMDT
               0,862    0,028    0,828      0,862    0,845      0,820    0,936     0,807     SMMA
               0,897    0,019    0,864      0,897    0,880      0,863    0,949     0,773     SMU
               0,788    0,019    0,852      0,788    0,819      0,795    0,917     0,711     URBS
               0,911    0,017    0,887      0,911    0,899      0,884    0,961     0,911     SMOP
               0,788    0,020    0,863      0,788    0,824      0,799    0,923     0,737     SMS
Weighted Avg.  0,846    0,022    0,847      0,846    0,845      0,824    0,933     0,775
```

uni9975-3 - resultado NBMultinomial - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayesMultinomial
Relation:     uni9975-3-weka.filters.unsupervised.attribute.Remove-R1
Instances:    37588
Attributes:   1004
              [list of attributes omitted]
Test mode:    10-fold cross-validation
```

uni9975-3 - resultado NBMultinomial - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
Correctly Classified Instances       34112               90.7524 %
Incorrectly Classified Instances      3476                9.2476 %
Kappa statistic                          0.8943
Mean absolute error                      0.0233
Root mean squared error                  0.1496
Relative absolute error                 10.6497 %
Root relative squared error             45.2492 %
Total Number of Instances            37588

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,966    0,002    0,985      0,966    0,976      0,973    0,994     0,986     FAS
                 0,886    0,031    0,802      0,886    0,842      0,820    0,971     0,802     SGM
                 0,830    0,016    0,881      0,830    0,855      0,835    0,966     0,884     SMDT
                 0,904    0,016    0,901      0,904    0,903      0,887    0,984     0,931     SMMA
                 0,933    0,020    0,867      0,933    0,899      0,885    0,985     0,917     SMU
                 0,932    0,006    0,952      0,932    0,942      0,934    0,993     0,967     URBS
                 0,919    0,004    0,970      0,919    0,944      0,936    0,983     0,951     SMOP
                 0,896    0,011    0,924      0,896    0,910      0,896    0,988     0,953     SMS
Weighted Avg.    0,908    0,013    0,910      0,908    0,908      0,895    0,983     0,924
```

APÊNDICE 11 – RESULTADOS OBTIDOS NA CLASSIFICAÇÃO COM BIGRAMAS – *DATASET* MAIOR

big9995-3 - resultado J48 - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     bi9995-3-weka.filters.unsupervised.attribute.Remove-R1
Instances:    37588
Attributes:   2118
              [list of attributes omitted]
Test mode:    10-fold cross-validation
```

big9995-3 - resultado J48 - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
Correctly Classified Instances       32129               85.4767 %
Incorrectly Classified Instances      5459               14.5233 %
Kappa statistic                          0.8338
Mean absolute error                      0.0477
Root mean squared error                  0.1567
Relative absolute error                 21.8196 %
Root relative squared error             47.3848 %
Total Number of Instances            37588

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,945    0,001    0,988      0,945    0,966      0,962    0,992     0,975     FAS
               0,850    0,020    0,855      0,850    0,852      0,832    0,969     0,871     SGM
               0,713    0,007    0,938      0,713    0,811      0,797    0,958     0,856     SMDT
               0,784    0,009    0,935      0,784    0,853      0,836    0,973     0,904     SMMA
               0,906    0,007    0,950      0,906    0,928      0,918    0,984     0,933     SMU
               0,788    0,005    0,953      0,788    0,863      0,851    0,979     0,907     URBS
               0,888    0,002    0,987      0,888    0,935      0,928    0,984     0,946     SMOP
               0,970    0,117    0,564      0,970    0,713      0,690    0,971     0,859     SMS
Weighted Avg.  0,855    0,022    0,892      0,855    0,862      0,849    0,976     0,905
```

big9995-3 - resultado NBMultinomial - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayesMultinomial
Relation:     bi9995-3-weka.filters.unsupervised.attribute.Remove-R1
Instances:    37588
Attributes:   2118
              [list of attributes omitted]
Test mode:    10-fold cross-validation
```

big9995-3 - resultado NBMultinomial - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
Correctly Classified Instances       33677               89.5951 %
Incorrectly Classified Instances      3911               10.4049 %
Kappa statistic                          0.881
Mean absolute error                      0.0266
Root mean squared error                  0.1509
Relative absolute error                 12.1703 %
Root relative squared error             45.6456 %
Total Number of Instances            37588

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,961    0,004    0,967      0,961    0,964      0,960    0,996     0,984     FAS
               0,882    0,025    0,828      0,882    0,854      0,834    0,974     0,839     SGM
               0,833    0,016    0,879      0,833    0,855      0,836    0,968     0,903     SMDT
               0,889    0,035    0,804      0,889    0,844      0,819    0,982     0,936     SMMA
               0,935    0,015    0,895      0,935    0,915      0,903    0,986     0,930     SMU
               0,893    0,006    0,952      0,893    0,922      0,912    0,991     0,959     URBS
               0,911    0,007    0,950      0,911    0,930      0,920    0,986     0,950     SMOP
               0,873    0,011    0,928      0,873    0,900      0,885    0,989     0,951     SMS
Weighted Avg.  0,896    0,015    0,899      0,896    0,897      0,882    0,984     0,931
```

APÊNDICE 12 – RESULTADOS OBTIDOS NA CLASSIFICAÇÃO COM BIGRAMAS – *DATASET* MENOR

big999-3 - resultado J48 - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     bi999-3-weka.filters.unsupervised.attribute.Remove-R1
Instances:    37588
Attributes:   886
              [list of attributes omitted]
Test mode:    10-fold cross-validation
```

big999-3 - resultado J48 - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
Correctly Classified Instances       31917               84.9127 %
Incorrectly Classified Instances      5671               15.0873 %
Kappa statistic                          0.8273
Mean absolute error                      0.0495
Root mean squared error                  0.1591
Relative absolute error                 22.6281 %
Root relative squared error             48.1293 %
Total Number of Instances            37588

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,933    0,001    0,990      0,933    0,961      0,957    0,991     0,970     FAS
                 0,847    0,020    0,855      0,847    0,851      0,831    0,969     0,870     SGM
                 0,705    0,007    0,933      0,705    0,803      0,789    0,957     0,853     SMDT
                 0,767    0,007    0,944      0,767    0,846      0,831    0,971     0,896     SMMA
                 0,909    0,008    0,939      0,909    0,924      0,914    0,984     0,931     SMU
                 0,786    0,006    0,948      0,786    0,859      0,847    0,978     0,906     URBS
                 0,879    0,001    0,990      0,879    0,931      0,924    0,983     0,942     SMOP
                 0,973    0,123    0,552      0,973    0,704      0,682    0,968     0,849     SMS
Weighted Avg.    0,849    0,023    0,890      0,849    0,857      0,843    0,975     0,901
```

big999-3 - resultado k-NN - IBk - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
=== Run information ===

Scheme:       weka.classifiers.lazy.IBk -K 1 -W 0 -A "weka.core.neighboursearch.LinearNNSearch -A \"weka.core.EuclideanDistance -R first-last\""
Relation:     bi999-3-weka.filters.unsupervised.attribute.Remove-R1
Instances:    37588
Attributes:   886
              [list of attributes omitted]
Test mode:    10-fold cross-validation
```

big999-3 - resultado k-NN - IBk - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
Correctly Classified Instances       31417               83.5825 %
Incorrectly Classified Instances      6171               16.4175 %
Kappa statistic                          0.8122
Mean absolute error                      0.0453
Root mean squared error                  0.1726
Relative absolute error                 20.7177 %
Root relative squared error             52.195  %
Total Number of Instances            37588

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,925    0,005    0,957      0,925    0,940      0,933    0,985     0,959     FAS
                 0,811    0,029    0,798      0,811    0,805      0,777    0,940     0,790     SGM
                 0,742    0,024    0,815      0,742    0,777      0,748    0,939     0,819     SMDT
                 0,834    0,052    0,717      0,834    0,771      0,735    0,962     0,881     SMMA
                 0,882    0,009    0,931      0,882    0,906      0,894    0,968     0,882     SMU
                 0,794    0,014    0,885      0,794    0,837      0,817    0,968     0,882     URBS
                 0,880    0,005    0,961      0,880    0,919      0,909    0,970     0,912     SMOP
                 0,827    0,050    0,719      0,827    0,769      0,733    0,969     0,868     SMS
Weighted Avg.    0,836    0,024    0,844      0,836    0,838      0,815    0,962     0,873
```

big999-3 - resultado NBMultinomial - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
=== Run information ===

Scheme:       weka.classifiers.bayes.NaiveBayesMultinomial
Relation:     bi999-3-weka.filters.unsupervised.attribute.Remove-R1
Instances:    37588
Attributes:   886
              [list of attributes omitted]
Test mode:    10-fold cross-validation
```

big999-3 - resultado NBMultinomial - Bloco de Notas

Arquivo Editar Formatar Exibir Ajuda

```
Correctly Classified Instances       32668               86.9107 %
Incorrectly Classified Instances      4920               13.0893 %
Kappa statistic                          0.8503
Mean absolute error                      0.0337
Root mean squared error                  0.1616
Relative absolute error                 15.4242 %
Root relative squared error             48.8655 %
Total Number of Instances            37588

=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,947    0,006    0,951      0,947    0,949      0,943    0,994     0,977     FAS
                 0,849    0,024    0,832      0,849    0,841      0,818    0,970     0,824     SGM
                 0,774    0,014    0,889      0,774    0,827      0,807    0,963     0,881     SMDT
                 0,884    0,063    0,692      0,884    0,776      0,743    0,978     0,920     SMMA
                 0,916    0,014    0,902      0,916    0,909      0,897    0,984     0,927     SMU
                 0,847    0,008    0,939      0,847    0,890      0,877    0,988     0,940     URBS
                 0,892    0,008    0,941      0,892    0,916      0,905    0,983     0,944     SMOP
                 0,851    0,015    0,901      0,851    0,876      0,857    0,987     0,935     SMS
Weighted Avg.    0,869    0,019    0,878      0,869    0,871      0,853    0,981     0,918
```

APÊNDICE 13 – AUTORIZAÇÃO PARA USO DOS DADOS, DOCUMENTOS E DO NOME DA PREFEITURA NA PESQUISA

Curitiba, 28 de agosto de 2019.

À
SECRETARIA MUNICIPAL DE GOVERNO MUNICIPAL
SUPERINTENDÊNCIA EXECUTIVA
A/C SR. AIRTON SOZZI JÚNIOR

Ref.: *Projeto de Mestrado da Universidade Federal do Paraná*

Prezado Senhor:

Venho pela presente solicitar a Vossa Senhoria cessão das informações expostas adiante com o propósito de colaborar para o desenvolvimento da dissertação de mestrado intitulada “Aplicação de métodos de aprendizado de máquina para classificação de demandas por serviços municipais em central de atendimento ao cidadão”, do Programa de Pós-Graduação em Gestão da Informação (PPGGI) da Universidade Federal do Paraná e sob a orientação do Prof. Dr. Ricardo Mendes Junior.

A referida dissertação visa primordialmente analisar a base de dados abertos da Central de Atendimento 156, por meio do emprego da tecnologia de inteligência artificial, com aplicação de aprendizado de máquina (*machine learning*). Para tanto solicito, nos termos da Lei 12.527/2011 - Lei de Acesso à Informação:

a) disponibilizar a base de dados abertos da Central de Atendimento 156, ano de exercício 2018, nos mesmos moldes da planilha mensal atualmente disponibilizada no Portal de Dados Abertos da Prefeitura de Curitiba (sítio internet);
b) disponibilizar a documentação referente ao histórico e aos processos da Central de Atendimento 156;

c) permissão para análise dos dados de atendimento da Central, durante o período de execução do mestrado, para fomento da base analítica do projeto de dissertação;
d) permissão para acompanhar o teleatendimento com vistas a compreender as etapas utilizadas pelos atendentes para classificar os chamados;
e) autorização para divulgar o nome da Prefeitura de Curitiba na dissertação e em demais publicações derivadas deste trabalho de mestrado.

As técnicas de aprendizado de máquina são capazes de identificar relacionamentos implícitos e reconhecer padrões em grandes quantidades de dados, reunindo informações que permitem a obtenção de previsões e estimativas, em apoio à tomada de decisão. No cenário dos dados da Central de Atendimento 156, a pretensão do estudo acadêmico recai no uso destas técnicas para identificar automaticamente a secretaria ou órgão responsável das demandas, a partir do treinamento de um algoritmo que se baseia na experiência e, a partir de exemplos (dados de 2018) consegue identificar padrões e antecipar eventos ou situações até então desconhecidas.

Assim sendo, o projeto prevê que a partir do texto livre, digitado pelo cidadão, sem a seleção prévia do serviço ofertado na lista do Sistema 156 *Web* ou do aplicativo *mobile* Curitiba 156, o algoritmo identifique automaticamente o ente interno administrativo responsável pela demanda.

De natureza igual, o trabalho objeto poderá auxiliar a Prefeitura na aplicação de métodos de inteligência artificial à classificação automática de demandas registradas pelos cidadãos, utilizando moderna tecnologia na tramitação inicial de protocolos, bem como para emissão de respostas automáticas, em se tratando de pedidos de informação, por exemplo, permitindo um sistema de atendimento com retorno imediato, *online* ao cidadão.

Na administração direta executiva a inteligência artificial tem se mostrado eficiente na regência tributária e, na área jurídica, tem-se adquirida habilidades na classificação célere e automática da jurisdição processual. O presente trabalho de mestrado tem a singela pretensão de contribuir e ajustar o cumprimento das tarefas da administração pública com a inteligência artificial, em se tratando do relacionamento com os cidadãos, consoante a finalidade fundamental do interesse público.

Fico à disposição para esclarecimentos adicionais e no aguardo de sua aprovação positiva, para este projeto que corrobora com a eficiência no atendimento dos anseios e necessidades da sociedade com o Poder Público.

Agradecendo a sua compreensão e colaboração, subscrevo-me.

Atenciosamente,

_____

Lucimara Wons.
Analista de Sistemas - Matrícula 81.184
Instituto de Pesquisa e Planejamento de Curitiba

Aprovação PMC/SGM

Carimbo, assinatura e data

Prefeitura Municipal de Curitiba **lwons@ippuc.org.br**

Fwd: Cessão Dados Abertos

De : Willian Jonderlan de Oliveira Belem <wbelem@sgm.curitiba.pr.gov.br> Qui, 05 de set de 2019 16:03
Assunto : Fwd: Cessão Dados Abertos 3 anexos
Para : Lucimara Wons <lwons@ippuc.org.br>

Oi,

Segue para conhecimento!

Willian Jonderlan

Gestor Central 156
SGM | Departamento de Programas e Projetos
(41) 3350-8727 | (41) 9639-8400

CURITIBA

Av. Cândido de Abreu, 817
Centro Cívico - (41) 3350-8484

De: "Juliana Midori de Carvalho Komiyama Catarino" <jmidori@smcs.curitiba.pr.gov.br>
Para: "Willian Jonderlan de Oliveira Belem" <wbelem@sgm.curitiba.pr.gov.br>
Cc: "Fabíola Maziero Pinheiro Sant'anna" <fmaziero@smcs.curitiba.pr.gov.br>
Enviadas: Quinta-feira, 5 de setembro de 2019 15:15:57
Assunto: Re: Cessão Dados Abertos

Oi Willan,

Não tem problema não, se os dados são abertos e vocês autorizaram o estudo, pode citar a Prefeitura de Curitiba. Sem problemas.

Pelo o que eu entendi o que ela quer é dizer qual é o "156" o qual ela está fazendo o estudo.

Beijos e obrigada

Juliana Midori

Superintendente
Secretaria Municipal da Comunicação Social
(41) 3350-8234

Av. Cândido de Abreu, 817
Centro Cívico - (41) 3350-8484
www.curitiba.pr.gov.br

CURITIBA

De: "Willian Jonderlan de Oliveira Belem" <wbelem@sgm.curitiba.pr.gov.br>
Para: "Juliana Midori de Carvalho Komiyama Catarino" <jmidori@smcs.curitiba.pr.gov.br>, "Fabíola Mazíero Pinheiro Sant'anna" <fmazíero@smcs.curitiba.pr.gov.br>
Enviadas: Quinta-feira, 5 de setembro de 2019 14:51:03
Assunto: Re: Cessão Dados Abertos

Boa tarde Juliana e Fabíola,

Recebemos um pedido da Lucimara Wons do IPPUC para realizar um Projeto de mestrado sobre a aplicação de tecnologia no sistema 156, o Sozzi já conversou com ela e autorizou, ficamos somente com dúvida sobre o assunto mencionado abaixo se existem algum problema mencionar que é da prefeitura de Curitiba esse sistema.

e) autorização para divulgar o nome da Prefeitura de Curitiba na dissertação e em demais publicações derivadas deste trabalho de mestrado.

Segue em anexo a solicitação completa.

Atenciosamente.

Willian Jonderlan

Gestor Central 156
SGM | Departamento de Programas e Projetos
(41) 3350-8727 | (41) 9639-8400

CURITIBA

Av. Cândido de Abreu, 817
Centro Cívico - (41) 3350-8484

De: "Lucimara Wons" <lwons@ippuc.org.br>
Para: "Willian" <wbelem@sgm.curitiba.pr.gov.br>
Enviadas: Sábado, 31 de agosto de 2019 19:48:34
Assunto: Cessão Dados Abertos

Agora sim, cessão escrito certo

Atenciosamente,

Lucimara Wons
DAF/Processos e Contratos
IPPUC Instituto de Pesquisa e Planejamento Urbano de Curitiba
+ 55 41 3250-1398

----- Em 31 de Ago de 2019, em 10:04, lucimara wons <lwons@ippuc.org.br> escreveu:

> Olá Willian
>
> Segue documento digital, caso precise.
>
> Muito obrigada pela seu apoio!!
>
> Atenciosamente,
>
> Lucimara Wons
> DAF/Processos e Contratos
> IPPUC Instituto de Pesquisa e Planejamento Urbano de Curitiba
> + 55 41 3250-1398